HUMBOLDT-UNIVERSITÄT ZU BERLIN

Dynamic regulation of co-transcriptional processes during neuronal maturation

DISSERTATION

zur Erlangung des akademischen Grades

bzw. Doctor of Philosophy

(Ph.D.)

eingereicht an der

Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

M.Sc. Ana Miguel Guterres Coelho Fernandes

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Dr. Ana Pombo
2. Prof. Dr. Leonie Ringrose
3. Prof. Dr. Shona Murphy

Tag der mündlichen Prüfung: 22.01.2020

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/ Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Declaration

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Berlin, ...

.....

Ana Miguel G. C. Fernandes

Abstract

Rpb1 is the largest subunit of RNA polymerase II (RNAPII) and it contains a long C-terminal domain (CTD) which is post-translationally modified at different stages of the transcription cycle. Coordinated phosphorylation of RNA polymerase II (RNAPII) CTD is essential for efficient coupling of nascent RNA synthesis with co-transcriptional RNA processing events. Circular RNAs (circRNAs) are a novel class of RNAs which are most abundant in neuronal cells. Their biogenesis remains ill understood, namely why the intron upstream to the circRNA is retained during transcription of circRNA exon(s) to enable back-splicing. Evidence suggests that deficient spliceosome recruitment can lead to circRNA formation.

To investigate the mechanisms that may be involved in deficient recognition and splicing of introns upstream of exons included in circRNAs, I mapped the chromatin occupancy of RNAPII phosphorylated forms, splicing factors and transcription regulators by ChIP-seq during mouse ESC differentiation to dopaminergic and spinal motor neurons. CircRNAs were detected throughout differentiation, peaking in differentiated neurons, as expected. I found that circRNAs are detected when their genes express the highest mRNA levels, and confirmed that circRNAs are most often produced from exon 2. Detailed investigation of RNAPII occupancy in mESCs showed that circRNA production is associated with lower abundance of promoter-proximal RNAPII-S7p and lower recruitment of the splicing machinery at the first exon-intron splice junction of circRNA-producing genes. Although RNAPII is recruited and initiates transcription with similar efficiency, the recruitment of promoter-proximal factors NELF and CDK9, and U1 snRNP is diminished at the promoters of circRNA-producing genes. Similar observations were found in both dopaminergic and spinal motor neurons, suggesting a common mechanism underlying circRNA formation. To mechanistically interfere with pausing mechanisms, I used siRNA-

mediated RNA interference to deplete NELF-A, a subunit of NELF complex, and discovered that NELF depletion was sufficient to increase the formation of circRNAs in mESCs. Our results implicate RNAPII regulation mechanisms in the formation of circRNAs. Finally, I propose a model for circRNA formation where RNAPII fast release from the promoter, possibly without sufficient phosphorylation of the S7p isoform, leads to altered spliceosome recruitment, ultimately favouring circRNA production over canonical splicing.

Zusammenfassung

Rpb1 ist die größte Untereinheit von RNA Polymerase II (RNAPII) und sie enthält eine lange C-terminale Domäne (CTD), welche an unterschiedlichen Zeitpunkten im Transkriptionszyklus post-translational modifiziert wird. Koordinierte Phosphorylierung der CTD von RNAPII ist essentiell für eine effiziente Kupplung von naszierender RNA Synthese und co-transkriptionalem RNA Prozessierens. Zirkuläre RNAs (circRNAs) sind eine neue Klasse von RNA Molekülen mit hoher Prävalenz in neuronalen Zelltypen. Die Biogenese von circRNAs ist noch ungeklärt, insbesondere die Frage warum das Intron upstream der circRNA während der Transkription des circRNA Exons zurückbehalten wird um Rück-Spleißen zu ermöglichen. Verschiede Belege suggerieren, dass unzulängliche Rekrutierung des Spleiceosoms zur circRNA Formation führen kann.

In dieser Arbeit untersuche ich die Mechanismen die zu Defekten in der Erkennung und des Spleißens des Introns upstream der circRNA führen. Mit diesem Ziel erfasste ich die genomweite Verteilung von chromatinassoziierter RNAPII mit verschiedenen Phosphorylierungen, sowie Spleißfaktoren und Transkriptionsreglern mittels ChIP-seq in neuronaler Differenzierung von murinen embryonalen Stammzellen zu dopaminergen und Motoneuronen. Während der gesamten Differenzierung, aber insbesondere in den differenzieren Neuronen, konnten circRNAs detektiert werden. Dabei werden circRNAs in Genen detektiert, wenn diese ihre höchsten Level an mRNA exprimieren, und die circRNAs entstehen in den meisten Fällen vom zweiten Exon des Gens. Eine detaillierte Untersuchung der RNAPII Verteilung im Stammzellgenom zeigt, dass die Produktion von circRNAs mit einer niedrigen Menge an Promoter-proximaler RNAPII-S7p und einer verminderten Rekrutierung des Spleiceosoms zur ersten Exon-Intron Spleißverbindung des circRNA-produzierenden Gens verbunden ist. Obwohl RNAPII mit ähnlicher Effizienz rekrutiert wird und Transkription initiiert, ist die Rekrutierung der Promoter-proximalen Faktoren NELF und CDK9,

sowie U1 snRNP, an den Promotern von circRNA-produzierenden Genen reduziert. Ähnliche Beobachtungen konnten in dopaminergen, sowie in Motoneuronen gefunden werden, was einen gemeinsamen Mechanismus hinter deren circRNA Formation suggeriert. Ich nutzte siRNA-mediierte RNA Interferenz um NELF-A auszuschalten, eine Untereinheit des NELF Komplexes, um mechanistisch in Promoter-Pausing Mechanismen einzugreifen. Hierbei entdeckte ich, dass das Entfernen von NELF ausreichte um die Formation von circRNAs in murinen embryonalen Stammzellen zu erhöhen. Unsere Ergebnisse implizieren, dass RNAPII Regulation eine Rolle spielt in der Formation von circRNAs. Schlussendlich formuliere ich ein Model zur Bildung von circRNAs, in dem die schnelle Freisetzung von RNAPII vom Promoter, potenziell ohne ausreichende Phosphorylierung der S7p Isoform, zu einer veränderten Rekrutierung des Spliceosoms führt, und damit zur bevorteilten Produktion von circRNAs gegenüber kanonischem Spleißen.

Acknowledgements

Saying that the PhD. is an adventure is little; it feels more like climbing Mount Everest…Ten times. It was a long journey with some difficult moments but also with immense joy. This experience definitely shaped me as a scientist and, most importantly, as a person.

First and foremost, my biggest thanks goes to Ana. I am deeply grateful for all the support, discussions, ideas, mentoring and the endless passion for science. You taught me how to be a scientist and that my data is as good as how I share it with the world.

I thank my co-supervisor Esteban Mazzoni, for supporting me throughout this journey, for all the advice, critical thinking, and for telling me to "let the data speak". I will never forget that lesson. I also thank MDC-NYU program for the amazing opportunity of experiencing science in Berlin and New York.

I also thank my PhD. committee member Uwe, for all the great feedback and support throughout these years.

A huge thanks to Sasha, for teaching me most of what I learned in the lab, all the support and advice, and for always pulling me up when I thought things would not work out. You were right, they did.

I thank Carmelo for welcoming me in the Pombo lab with a big hug and guiding me through the early steps of my project and throughout.

I thank Markus and Tiago for initiating me in the arts of bioinformatics and giving me the courage to believe that one day I would make my scripts work.

A big thanks to Iza for forcing me to take breaks, all the talks, laughs, memes, puppies, and other hidden things (you know what I mean).

I thank the former and present members of Beautiful People: Marta, Elena, Giulia, Mariano, Kelly, João, Rob, Christophe, Dominik, Tom, Luna, Sílvia, Ehsan, Gesa, Rieke, Ibai, Warren and Jenny. Thank you for all the talks, laughs, balloons, sparkles, and the great times together.

A huge thanks to my dear friends João, Kamila, Andrea, Neşe and Can. These crazy years would not have been as amazing without you. You walked with me through my worst and my best moments, and showed me that even if you are far from home you can still find your family.

I thank my beautiful friends from Portugal (and now kind of spread all over the globe): Andreia, Bia (do Hip Hop), Bia (de Fátima), Bota, Maria e Joana Pê. I am very grateful for growing with you since the first year of university and for still sharing this wonderful friendship despite being far apart.

Lastly, a huge thanks to my family, especially my mom and dad, for the unconditional support through all the years. Thank you for believing in me, for stimulating my curiosity and being very patient, even when a four-year old is asking you how washing machines work. Thanks for letting me fly.

Por último, um imenso obrigada à minha família, especialmente à minha mãe e ao meu pai. Obrigada pelo suporte incondicional durante todos estes anos. Obrigada por acreditarem em mim, por sempre estimularem a minha curiosidade e por toda a paciência, mesmo quando uma criança de quatro anos vos pergunta como functionam as máquinas de lavar. Obrigada por me deixarem voar.

Table of Contents

Index of Figures

Index of Tables

Abbreviations

3'ss	3' splice site
5'ss	5' splice site
ADAR1	Adenosine deaminase acting on RNA 1
AID	Auxin-inducible degron
AS	Alternative splicing
BCP	Bayesian Change Point
BPS	Branch point signal
CARM1	Coactivator-associated arginine methyltransferase 1
CBC	Cap binding complex
CF1	Cleavage factor 1
CF2	Cleavage factor 2
ChIP	Chromatin immunoprecipitation
circ-	Genes never producing circRNAs in dopaminergic and spinal differentiations
circ+	Genes producing circRNAs at a given time-point
circRNAs	Circular RNAs
$CircRP_{100}M$	CircRNA expression per gene
CLIP	Crosslinking immunoprecipitation
CPSF	Cleavage and polyadenylation specificity factor
CstF	Cleavage stimulatory factor
CTD	C-terminal domain
D. melanogaster	Drosophila melanogaster
DHX9	DexH-Box Helicase 9
Dig	Digoxigenin antibody
DRB	5,6-dichloro-1-β -D-ribofuranosylbenzimidazole
DSIF	DRB-sensitivity inducing factor
dsRNA	Double stranded RNA
E/ISE	Exonic and intronic splicing enhancer
E/ISS	Exonic and intronic splicing silencer
EBs	Embryoid bodies
EpiSCs	EpiStem cells
FUS	Fused in Sarcoma
GAF	GAGA factor
GO	Gene Ontology
GTFs	General transcription factors
H3K36me3	H3 on Lys 36
H3K4me3	H3 at lysine 4
HB9	Motor Neuron And Pancreas Homeobox 1
hnRNPs	heterogeneous ribonucleoproteins
K7ac	RNA Polymerase II lysine acetylation

K7me2	RNA Polymerase II lysine dimethylation
KD	Knockdown
Mbl	Muscleblind
me7Gppp	Methylated guanine nucleoside
mESCs	Mouse ES cells
miRNA	microRNA
mRNA	Messenger mRNA
NELF	Negative elongation factor
NIL	Ngn2, Isl1, Lhx3 transcription factors
NPCs	Neuronal progenitor cells
O-GlcNAc	O-linked N-acetyl-glucosamine
OGA	N-acetylglucosamidase
OGT	O-GlcNAc transferase
PAP	Poly(A) polymerase
PAS	Poly(A) signal
PIC	Pre-initiation complex
PPT	Polypyrimidine tract
pre-EJC	pre-Exon Junction Complex
pre-mRNAs	Precursor mRNAs
PTBP1	Polypyrimidine tract-binding protein 1
QKI	Quaking
RBPs	RNA binding proteins
RNAP	RNA polymerase
RNAPII	RNA polymerase II
RPRD	Regulator of pre-mRNA-domain-containing
S2p	RNA Polymerase II - Serine 2 phosphorylation
S5p	RNA Polymerase II - Serine 5 phosphorylation
S7p	RNA Polymerase II - Serine 7 phosphorylation
SCR	Scrambled
SF1	Splicing factor 1
siRNA	Small interfering RNA
SMN	Survival of Motor Neuron
snRNPs	Small nuclear ribonucleoproteins
SR	Serine/Arginine proteins
SRRM4	Serine/Arginine repetitive matrix protein 4
T4p	RNA Polymerase II - Threonine 4 phosphorylation
TAFs	TBP-associated factors
TBP	TATA-box binding protein
TES	Transcription end site
TFs	Transcription factors
TH	Tyrosine Hydroxylase
TPM	Transcript Reads per Million
TSS	Transcription start site
TUBB3	β-tubulin III

U2AF	U2 Auxiliary factor
Y1p	RNA Polymerase II - Tyrosine 1 phosphorylation

Notes to the reader

Contribution to this thesis:

From Pombo group: Carmelo Ferrai provided RNA samples and produced total RNA-seq library production for biological replicate 1 from the dopaminergic neuron differentiation (mESCs to day 30). Giulia Caglio mapped the total RNA-seq datasets for biological replicate 1 from the dopaminergic neuron differentiation (mESCs to day 30) after sequencing. Alexander Kukalev provided RNA samples for biological replicate 2 from the dopaminergic neuron differentiation (days 16 and 30), contributed to the production of ChIP-seq libraries (S5p, Dig and Mock in mESCs) and taught me ChIP-seq quality controls, as well as designed the NELF knockdown experiments. Izabela Harabula provided chromatin samples for ChIP of U1C and NELF-E in dopaminergic neurons day 16. Markus Schueler contributed to the calculation of the number of circRNAs per gene. Tiago Rito devised the strategy for normalization of circularized reads per gene, with gene filtering analyses to define circ+ and circ- genes, guided me through ChIP-seq analyses and provided the script for calculating the coverage of ChIP-seq enrichment. Christoph Thieme, Warren Winick-Ng and Dominik Szabo advised on the statistical analyses of NELF knockdown experiments.

From Nikolaus Rajewsky group, MDC: Petar Glažar performed circRNA identification together with a list with matched linear transcripts and performed corresponding quality controls for the dopaminergic and spinal motor neuron differentiation and NELF knockdown experiments.

From Esteban Mazzoni group, NYU: Silvia Velasco taught me the spinal motor neuron differentiation and provided the protocol for ChIP in embryoid bodies which I later adapted for RNAPII ChIP and Disi An advised on spinal motor neuron re-plating.

Permissions to reuse figures from published papers are provided in Appendix.

"Can we climb this mountain
I don't know
Higher now than ever before
I know we can make it if we take it slow
Let's take it easy
Easy now, watch it go

We're burning down the highway skyline
On the back of a hurricane that started turning
When you were young"

The Killers

Part I

Introduction

1 Introduction

Higher organisms have evolved numerous strategies to cope with developmental and environmental challenges in a balanced manner. To respond to both internal and external stimuli, organisms must activate complex networks to tightly regulate transcriptional output and ultimately cellular function. This is achieved through precise modulation of transcription and RNA processing.

Gene expression is a complex process, shaped by the relationship between the transcription machinery, the chromatin environment, and RNA maturation. A key player in integrating different levels of gene expression is RNA polymerase II (RNAPII). Its largest subunit, RPB1, contains a long, highly repetitive C-terminal domain (CTD) that is heavily post-translationally modified. CTD modifications change dynamically through the transcription cycle and promote the recruitment of different machineries, such as chromatin modifiers and RNA processing factors, which facilitates co-transcriptional RNA maturation.

This thesis aims to investigate the interplay between the CTD modifications of RNAPII and differential RNA output, focusing on the role of RNAPII modifications in circular RNA (circRNA) formation during neuronal maturation. In this introduction, I summarise transcription, splicing, and how the modifications of RNAPII's CTD mediate both processes. I also describe the function and biogenesis of circular RNAs.

1.1 Transcription and RNAPII regulation

Gene expression starts when DNA-dependent RNA polymerases (RNAPs) read genetic information encoded in the DNA sequence and transform it into RNA. In eukaryotes, RNAPs evolved into three distinct complexes that transcribe specific gene groups: RNAPI, RNAPII and RNAPIII. RNAPI transcribes ribosomal RNA genes in the nucleolus, which can comprise up to 80% of the cellular RNA.

RNAPIII transcribes mainly transfer RNAs, as well as other small RNAs, such as 7SK RNA and 5S rRNA. In this thesis, I focus on RNAPII, which transcribes all protein-coding genes and many structural and non-coding RNAs (Khatter, Vorländer, and Müller 2017).

1.1.1 RNAPII features and the transcription cycle

RNAPII is a multimeric protein complex composed of 12 subunits. Its catalytic subunit, RPB1, contains a long CTD, which is highly conserved, structurally disordered and composed of multiple repeats with the canonical sequence Y^{1}-S^{2}-P^{3}-T^{4}-S^{5}-P^{6}-S^{7}. The number of repeats ranges from 26 in yeast to 52 in mammals and correlates with organism complexity (Eick and Geyer 2013). Although the CTD is not necessary for RNAPII catalytic activity, its deletion or extensive truncation is lethal, showing that the CTD is essential for cell survival (Bartolomei et al. 1988; Thompson et al. 1993). In mammals, the proximal part of the CTD is highly conserved; however, the distal part of the CTD shows significant changes from the consensus sequence most often at position 7 (Eick and Geyer 2013). All heptapeptide residues can be post-translationally modified, which alter the functionality and structure of the CTD and in turn influence its interaction partners. The CTD can be subjected to different modifications, such as phosphorylation, methylation, acetylation, glycosylation and isomerization. The best studied modifications are phosphorylation of Ser2, Ser5 and Ser7 (S2p, S5p, S7p, respectively, **Fig. 1.1A**). Mapping these modifications by chromatin immunoprecipitation (ChIP) followed by high throughput sequencing (ChIP-seq) shows that both S5p and S7p peak at the transcription start site (TSS), whereas S2p gradually increases until it peaks at the transcription end site (TES) (Brookes and Pombo 2009; Zaborowska, Egloff, and Murphy 2016). The extent to which the CTD is modified and which combinations are present is still under investigation. Two recent studies combined genetic manipulation and mass spectrometry to explore the relative amounts and spatial distribution of the different CTD modifications in yeast and mammals (Schuller et al. 2016; Suh et al. 2016). The CTD tends to be uniformly modified and most often once per repeat, although

double modifications per repeat were also found. About 75% of the repeats were phosphorylated on S5p and S2p, with tyrosine 1 phosphorylation (T1p), threonine 4 phosphorylation (T4p), and S7p accounting for the remaining 15%. Importantly, different modifications have been linked to different stages of the transcription cycle (Brookes and Pombo 2009; Zaborowska, Egloff, and Murphy 2016), where RNAPII CTD acts as a docking platform for different factors essential for nascent RNA processing and shaping the chromatin environment (**Fig. 1.1B**) (David et al. 2011; McCracken, Fong, Rosonina, et al. 1997; McCracken, Fong, Yankulov, et al. 1997; Morris and Greenleaf 2000; Ryan et al. 2002; Rosonina and Blencowe 2004). For example, depleting CTD phosphorylation by mutating all Ser2, Ser5 or Ser7 residues to alanine or glutamate is lethal in yeast and mammalian cells, showing that mutations in these residues are not tolerated (West and Corden 1995; Zhang et al. 2012). The CTD modifications are placed by kinases and removed by phosphatases and their combined and timely action during the transcription cycle are essential for proper CTD function. The transcription cycle of protein-coding genes is composed of the following stages: initiation, promoter-proximal pausing, elongation and termination. For this, RNAPII is first recruited to the promoter, then transcribes through the coding region and terminates transcription at the end of the gene (**Fig. 1.1B**).

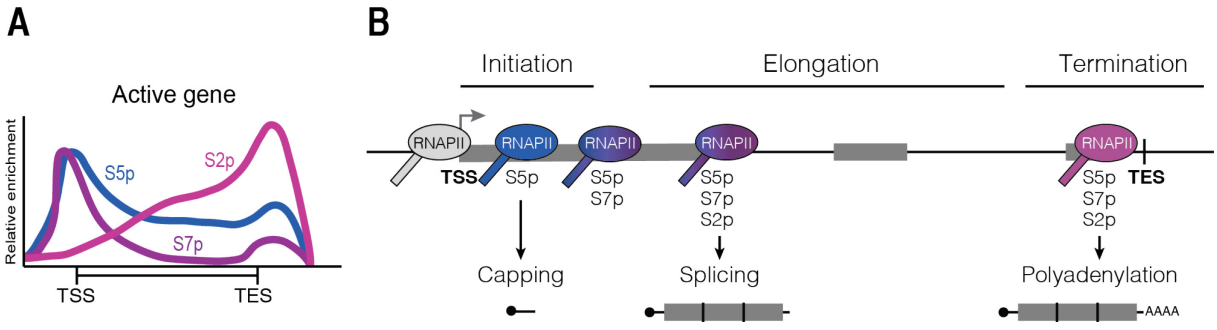

Figure 1.1 RNAPII S2p, S5p and S7p change dynamically during the transcription cycle.
A) Schematic representation of RNAPII S2p, S5p and S7p relative enrichment on chromatin of active genes. **B)** Diagram depicting connections between transcription cycle stages, RNAPII modifications, and co-transcriptional RNA processing. TSS – Transcription start site. TES – Transcription end site.

1.1.2 Initiation and S5p

For RNAPII to initiate, several transcription factors (TFs) and protein complexes are recruited to open chromatin regions (reviewed in detail in (Sainsbury, Bernecky, and Cramer 2015) and (Grunberg and Hahn 2013). Firstly, general transcription factors (GTFs) TFIIA, B and D are recruited to the promoter, forming the core initiation complex. TFIID is composed of TATA-box binding protein (TBP), required for basal transcription, and several TBP-associated factors (TAFs), which are promoter specific. After the assembly of the core initiation complex, TFIIF and unphosphorylated RNAPII are recruited to the promoter, followed by TFIIE and TFIIH, forming the closed pre-initiation complex (PIC). Upon ATP hydrolysis, XPB, the helicase subunit of TFIIH, melts the DNA strands and creates a transcription bubble, called the open PIC. Unphosphorylated CTD has a negative impact on transcription and promotes RNAPII stalling at the PIC, because both GTFs and the Mediator complex have high affinity for unphosphorylated RNAPII (Maxon, Goodrich, and Tjian 1994; Myers et al. 1998).

The Mediator is a multiprotein complex that is thought to play a central role in activating gene expression by integrating enhancer-promoter contacts (Allen and Taatjes 2015). Finally, CDK7, the kinase subunit of TFIIH, phosphorylates RNAPII CTD on Ser5, which decreases its affinity to the PIC, the mediator is evicted from the PIC, and RNAPII initiates transcription (Sogaard and Svejstrup 2007; Wong, Jin, and Struhl 2014). The Mediator further contributes to RNAPII release from the PIC by phosphorylating RNAPII CTD on Ser5 through one of its associated subunits, CDK8 (Jeronimo and Robert 2017).

RNAPII S5p role in transcription initiation

RNAPII S5p is most enriched at the TSS and marks transcription initiation. Ser5 is thought to be phosphorylated at active genes mostly by CDK7 (Akhtar et al. 2009), although CDK8, CDK9, CDK12, CDK13 and DYRK1A have been suggested to phosphorylate Ser5 based mostly on *in vitro* studies (Bosken et al. 2014; Czudnochowski, Bosken, and Geyer 2012; Di Vona et al. 2015; Greifenberg et al. 2016). Dephosphorylation of Ser5 residues depends on several phosphatases.

Ssu72 phosphatase dephosphorylates both S5p and S7p and is particularly important in the transition between initiation and elongation (Ganem et al. 2003; Krishnamurthy et al. 2004). RPAP2 and its yeast homolog Rtr1 also dephosphorylate S5p (Egloff et al. 2012; Mosley et al. 2009). Additionally, SCP1, part of the small phosphatases family, was reported to mediate S5p dephosphorylation and promote gene silencing of neuronal genes in non-neuronal cells (Eick and Geyer 2013).

S5p is intimately linked with early co-transcriptional events, for example capping and splicing. Capping consists of the addition of a methylated guanine nucleoside (me7Gppp) to the 5' end of the nascent RNA by the capping enzyme as soon as it exits RNAPII, when it is ~20 nucleotides long. This prevents the degradation of nascent RNA by 5'exonucleases and facilitates messenger mRNA (mRNA) translation. The capping enzyme was shown to directly bind to S5p (Fabrega et al. 2003) and S5p is essential for the co-transcriptional recruitment of the capping machinery to nascent RNA (Cho et al. 1997; Ghosh, Shuman, and Lima 2011; McCracken, Fong, Rosonina, et al. 1997; Schwer and Shuman 2011). For example, mutating S5 residues to alanines prevents recruitment of the capping machinery and is lethal in yeast and mammals; however, lethality can be overcome by tethering the capping enzyme to the nascent RNA (Schwer and Shuman 2011). Several studies also suggest that S5p helps recruit and interacts with spliceosome components (Harlen et al. 2016; Nojima et al. 2018). Finally, the state of chromatin modifications at histone tails is also is also shaped by S5p. For example, Set1 methyltransferase, which methylates histone H3 at lysine 4 (H3K4me3) and is a hallmark of active genes, binds to S5p (Kim and Buratowski 2009; Ng et al. 2003).

S5p modification, together with H3K4me3, are also found at Polycomb-repressed genes, whose promoters are marked by repressive chromatin marks (H3K27me3 and H2Aub1). At these genes, RNAPII is displays a "poised state", in which its CTD is phosphorylated on S5, but not on Ser2 or Ser7 (Brookes et al. 2012; Stock et al. 2007). Genes co-occupied by RNAPII and Polycomb complexes have been

described in mouse ES cells (mESCs) and throughout neuronal differentiation (Brookes et al. 2012; Ferrai et al. 2017). Finally, it was shown that S5p is modified by a different kinase, ERK1/2, at polycomb-repressed genes in mESCs (Tee et al. 2014).

Role of other RNAPII modifications in transcription initiation

Another RNAPII modification which may play a role in transcription initiation is Y1p. In mammals, Y1p is most enriched at the TSS, then reduced in the gene body, and enriched again at the TES but at a lower level (Descostes et al. 2014). Y1p was described to be associated with antisense promoter transcription and active enhancers (Descostes et al. 2014), but its function remains unclear. In yeast, Y1p is gradually enriched along the gene body and peaks at the TES (Mayer et al. 2012), suggesting that it plays a distinct role in this organism. Y1p is thought to be placed by c-ABL kinase in mammalian cells (Baskaran, Dahmus, and Wang 1993; Baskaran, Chiang, and Wang 1996), and is removed by Rtr1 and Glc7 in yeast (Hsu et al. 2014; Schreieck et al. 2014).

Finally, CTD glycosylation consists of the addition of O-linked N-acetyl-glucosamine (O-GlcNAc) is added to serine and threonine residues by O-GlcNAc transferase (OGT), is thought to modulate the RNAPII recruitment and initiation (Lewis, Burlingame, and Myers 2016). For example, OGT is found enriched at promoters, interacts with GTFs and is necessary for RNAPII recruitment to the promoter (Comer and Hart 2001; Kelly, Dahmus, and Hart 1993; Ranuncolo et al. 2012). Moreover, glycosylated RNAPII is found at the PIC and glycosylation prevents CTD phosphorylation (Comer and Hart 2001; Kelly, Dahmus, and Hart 1993), suggesting that glycosylation may be an extra regulatory layer of upstream of transcription initiation.

1.1.3 Promoter-proximal pausing and S7p

After release from the PIC, RNAPII is phosphorylated on Ser7 and transcribes about 20-60 nucleotides downstream of the TSS, where it remains stably associated with the nascent RNA, a stage of the transcription cycle called "promoter-proximal pausing". This process is mainly mediated through the binding of Negative elongation factor (NELF) and DRB-sensitivity inducing factor (DSIF) complexes to RNAPII (Ni et al. 2008; Wu et al. 2003). Promoter-proximal pausing is reflected by the accumulation of RNAPII at promoters of many genes, modified by both S5p and S7p. Further signaling is necessary to release RNAPII into productive elongation: the PTEF-b complex is recruited to the promoter-proximal pausing site and its kinase subunit CDK9 phosphorylates NELF, DSIF and RNAPII on Ser2. Consequently, NELF is released from the RNAPII complex, DSIF becomes a positive elongation factor and RNAPII transitions into productive elongation (reviewed in (Adelman and Lis 2012; Chen, Smith, and Shilatifard 2018). These regulatory steps are depicted in **Fig. 1.2**.

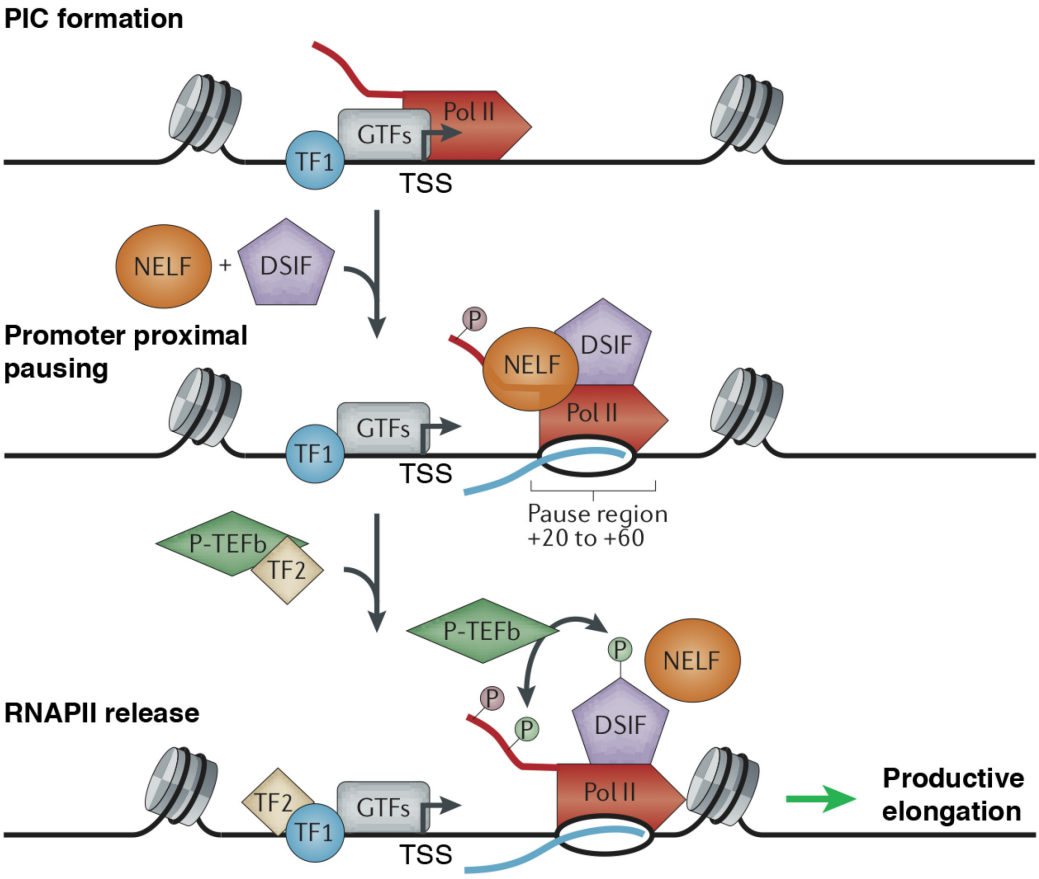

Figure 1.2 Schematic representation of promoter-proximal pausing establishment and release of RNAPII.
DNA is represented as a black line and nascent RNA is the blue line. Sequence-specific transcription factors (TF1) recruit GTFs and RNAPII (Pol II in red) to form the PIC. After RNAPII transcribed ~20 to 60 bp it is paused by the action of NELF and DSIF and promoter-proximal pausing is established. Release of RNAPII is triggered by the recruitment of P-TEFb kinase, directly or indirectly by another putative transcription factor (TF2). P-TEFb phosphorylates NELF, DSIF and RNAPII S2 and RNAPII is released into productive elongation (Adelman and Lis 2012).

Function of promoter-proximal pausing

Promoter-proximal pausing of RNAPII was initially described by the Lis lab at the heat shock gene (*Hsp70*) in *Drosophila melanogaster (D. melanogaster)*, but has since then been accepted as a widespread phenomenon in metazoans (Adelman and Lis 2012). Several studies have shown that genes with high accumulation of RNAPII at promoters compared to gene bodies, termed "paused genes", are enriched for fast response signaling processes, such as cell proliferation, stress response or development (Adelman et al. 2009; Aida et al. 2006; Henriques et al. 2013; Zeitlinger et al. 2007). This led to the proposal that RNAPII is present at the

promoters of these genes to allow quick activation in response to cellular stimuli. Nevertheless, not all rapidly induced genes are paused, and strongly paused genes are not necessarily highly inducible (Kininis et al. 2007; Lin et al. 2011). RNAPII accumulation was also suggested to help maintaining open chromatin regions at paused promoters (Mavrich et al. 2008; Ozsolak et al. 2007; Schones et al. 2008).

Promoter-proximal pausing is also thought as an important checkpoint in the transcription cycle, where either RNAPII terminates transcription or is released into productive elongation with the appropriate CTD modifications to recruit processing factors and chromatin modifiers (Adelman and Lis 2012). Stalling of RNAPII at the promoter-proximal pause site is thought to help recruit processing machineries to chromatin and nascent RNA. Supporting this hypothesis, DSIF complex was found to interact with the 5' capping machinery (Mandal et al. 2004; Pei and Shuman 2002; Wen and Shatkin 1999). Recent studies also suggest that promoter-proximal pausing may be the limiting step in the transcription cycle instead of RNAPII recruitment (Bartman et al. 2019; Gressel et al. 2017; Shao and Zeitlinger 2017).

Finally, the promoter release of paused RNAPII may be modulated by promoter enhancer contacts and proteins which are also involved in genome architecture. For example, PAF1, which regulates promoter-proximal pausing, also mediates promoter-enhancer contacts (Chen et al. 2017; Chen et al. 2015; Yu et al. 2015). In *Drosophila*, cohesin-bound genes have high transcription levels and are more paused. Depletion of cohesin decreased transcription and increased promoter-proximal pausing, suggesting that cohesin could facilitate RNAPII release into productive elongation (Schaaf et al. 2013). CTCF was also shown to modulate recruitment of NELF, SPT5 and CDK9 at the promoter-proximal pausing site of *c-myc* gene in HeLa cells (Laitem et al. 2015), suggesting that CTCF could modulate the promoter-proximal pausing of genes that depend on CTCF binding. To conclude, promoter-proximal pausing is a key step in modulating the transcription cycle and gene expression.

Regulation of promoter-proximal pausing

Establishment of RNAPII pausing near promoters and subsequent release is intricately regulated and the location and duration of RNAPII pausing at promoters depends on various factors, such as DNA sequence and regulatory proteins. Specific DNA features were found associated with increased levels of promoter-proximal pausing, such as GC-richness, lack of TATA-box or formation of R-loops, which are DNA-RNA hybrids that tend to form on GC-rich sequences (Chen et al. 2017; Core, Waterfall, and Lis 2008; Day et al. 2016; Skourti-Stathaki and Proudfoot 2014). Additionally, RNAPII pausing at promoters is tightly modulated by several protein complexes, such as DSIF and NELF (Adelman and Lis 2012; Chen, Smith, and Shilatifard 2018). The DSIF complex is composed of SPT4 and SPT5. SPT5 directly contacts the nascent RNA as it exits RNAPII (Bernecky, Plitzko, and Cramer 2017; Ehara et al. 2017; Qiu and Gilmour 2017). DSIF is recruited to the RNAPII complex upon TFIIE phosphorylation by CDK7 and subsequent TFIIE eviction from the PIC (Larochelle et al. 2012). Furthermore, SPT5 has a repetitive CTD, similar to RNAPII's CTD, which contains 11 nonapeptide tandem repeats and is mostly phosphorylated by CDK9, but also by CDK7. Upon CDK9 phosphorylation, DSIF becomes a positive elongation factor which remains associated with RNAPII until it terminates transcription (Adelman and Lis 2012). The second promoter-proximal pausing complex, NELF, is composed of 4 subunits: A, B and C (or D, its isoform) and E, its catalytic subunit. All four NELF subunits interact with nascent RNA, which suggests that NELF may help stabilize paused RNAPII complexes and prevent premature transcription termination (Vos et al. 2016; Yamaguchi et al. 2002). Importantly, NELF is a potent inhibitor of transcription in the presence of DSIF *in vitro* (Yamaguchi et al. 1999) and is released from paused RNAPII complexes upon phosphorylation by CDK9 (Nechaev and Adelman 2011; Peterlin and Price 2006; Sims, Belotserkovskaya, and Reinberg 2004). Finally, other protein complexes were found to enhance RNAPII pausing at promoters. For example, GDOWN1 prevents TFIIF recruitment and RNAPII premature termination (Cheng et al. 2012; DeLaney and Luse 2016; Workman and Roeder 1987; Jishage et al. 2012; Jishage

et al. 2018). Additionally, the PAF1 complex mediates RNAPII release into productive elongation (Chen et al. 2015; Xu et al. 2017; Yu et al. 2015; Vos et al. 2018).

As mentioned above, release of RNAPII from promoter-proximal pausing into productive elongation is also highly regulated, where the PTEF-b complex plays a central role. The PTEF-b complex is composed of a cyclin subunit (T1, T2, or K) and a kinase subunit, CDK9 (Adelman and Lis 2012). To release RNAPII from pausing, PTEF-b is recruited to paused RNAPII complexes and CDK9 phosphorylates NELF, SPT5 and RNAPII on Ser2 (Nechaev and Adelman 2011; Peterlin and Price 2006; Sims, Belotserkovskaya, and Reinberg 2004). Chemical inhibition of CDK9 by flavopiridol or 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) strongly prevents the release of RNAPII from the promoter into productive elongation (Chao et al. 2000; Peng et al. 1998; Laitem et al. 2015). Finally, RNAPII CTD modifications, namely S7p, also play an important role in the transition from initiation to elongation.

Promoter-proximal pausing is further regulated by the recruitment of PTEF-b complex to active genes, which in turn associates with distinct complexes that modulate its activity. Most PTEF-b complexes in the cell are associated with the 7SK ribonucleoprotein complex, which sequesters and inactivates PTEF-b (Nguyen et al. 2001; Yang et al. 2001), a process also mediated by the binding of HEXIM1/2 to PTEF-b (Michels et al. 2003). Release of PTEF-b from repression depends on KAP1/TRIM28 and association of splicing factors SRSF1/2 with 7SK RNA on chromatin (Ji et al. 2013; McNamara et al. 2016). Once released from the 7SK complex, PTEF-b may associate with two other protein complexes, the SEC and BRD4, where PTEF-b is active and able to trigger RNAPII promoter release (Jang et al. 2005; Lin et al. 2011; Luo et al. 2012; Yang et al. 2005). Both of these complexes have the ability to phosphorylate RNAPII on Ser2, but it remains unclear how they work together. The SEC contains multiple Lys-rich leukemia proteins and mixed lineage leukemia translocation partners which are often together with PTEF-b and are required for quick induction of transcription

(reviewed in detail in (Luo, Lin, and Shilatifard 2012)). BRD4 recruits PTEF-b to promoters and triggers RNAPII release into productive elongation (Bisgrove et al. 2007; Jang et al. 2005; Yang et al. 2005).

RNAPII S7p role in the transition to elongation

Of all the serine residues, S7p is so far the least understood. It has similar pattern to S5p on actively transcribed genes (Chapman et al. 2007; Egloff et al. 2007). It is placed by CDK7 kinase and removed by Ssu72 phosphatase, which also removes S5p (Akhtar et al. 2009; Jeronimo, Bataille, and Robert 2013; Glover-Cutter et al. 2009; Kim et al. 2009; Zhang et al. 2012). However, CDK7 may not be the only modifier of Ser7, as CDK9 and CDK12 seem to also phosphorylate Ser7 *in vitro* (Glover-Cutter et al. 2009; Bosken et al. 2014).

One of the best described roles of S7p is in snRNA gene expression, where it helps recruit the Integrator complex and RPAP2 (Egloff et al. 2007; Egloff et al. 2012). Furthermore, mutations of Ser7 residues to alanine causes expression defects of these genes (Egloff et al. 2007). S7p was also shown to participate in the suppression of cryptic transcription in yeast (Tietjen et al. 2010). Finally, S7p is thought to promote RNAPII release into productive elongation, as S7p presence primes RNAPII CTD for PTEF-b recognition and subsequent phosphorylation of Ser2 by CDK9 (Czudnochowski, Bosken, and Geyer 2012; St Amour et al. 2012; Viladevall et al. 2009). To summarize, S7p appears to be intimately linked with regulatory networks at both transcription initiation and the transition into productive elongation.

Role of lysine 7 modifications in the transition between initiation and elongation

In mammals, the distal part of the CTD has 8 lysine residues at the position 7 (K7). The number of lysine residues appears to correlate with organism complexity: most vertebrates have 8, *D. melanogaster* has 3, *Caenorhabditis elegans (C. elegans)* has 1 and no K7 residues are found in yeast (Dias et al. 2015). The first modification described in K7 residues was ubiquitination by ubiquitin E3 ligase WWP2, which causes proteosomal degradation of the RPB1 subunit in mESCs (Li

et al. 2007). K7 residues can also be acetylated (K7ac) *in vitro* and *in vivo* (Schroder et al. 2013). K7ac is enriched at the TSS of active genes and is placed by P300 kinase (Schroder et al. 2013). It is also implicated in transcription initiation and elongation: mutating K7 residues or inhibiting P300 prevents or reduced induction of EGF responsive genes *Egr2* and *c-Fos* (Schroder et al. 2013). A recent study in HEK293 cells has shown that the RPRD proteins specifically interact with K7ac and promote S5p dephosphorylation by a RPRD-associated phosphatase, thereby providing an additional regulatory step between initiation and early elongation in vertebrates which would contribute to precise release of RNAPII into productive elongation (Ali et al. 2019). Finally, K7 residues can be mono- or dimethylated (K7me1, K7me2) *in vivo* (Dias et al. 2015; Voss et al. 2015) and mark early stages of transcription initiation (Dias et al. 2015). K7me1 and K7me2 are enriched at the TSS of active genes similarly to K7ac, though K7me is strictly located at the TSS and K7ac extends into the gene body (Dias et al. 2015). The enzyme responsible for K7 methylation and potential readers are still unknown. The presence of methylation is not compatible with acetylation of the same K7 residue; nevertheless, different repeats on the same CTD may be methylated or acetylated. Remarkably, K7me1 and K7me2 are highly correlated with initiating and early elongating forms of RNAPII (S5p and S7p), whereas K7ac is positively correlated with elongating RNAPII (S2p) and mRNA levels. Finally, Dias and colleagues also found that the balance between K7me2 and K7ac promoter levels at each gene (K7me2/K7ac ratio) may be important to fine-tune gene expression (Dias et al. 2015).

1.1.4 Elongation and S2p

In contrast to RNAPII S5p and S7p which are most enriched at the promoter, S2p gradually increases as RNAPII transcribes through the coding region, marking the elongating form of RNAPII (Komarnitsky, Cho, and Buratowski 2000). As previously discussed, the enzyme mostly responsible for Ser2 phosphorylation is CDK9. However, Ser2 can also be modified by CDK12, CDK13 and DYRK1A *in vitro*, by CDK11 *in vivo*, and by BRD4 both *in vitro and in vivo* (Pak et al. 2015;

Bosken et al. 2014; Czudnochowski, Bosken, and Geyer 2012; Devaiah et al. 2012; Greifenberg et al. 2016; Karagiannis and Balasubramanian 2007). Ser2 and Ser5 share many of its modifiers which points towards an intimate cross-talk between transcription initiation and elongation. S2p levels at the TSS are regulated by the Fused in Sarcoma (FUS) protein which binds to the CTD and prevents Ser2 phosphorylation by CDK9 and CDK12 (Schwartz et al. 2012). Finally, S2p is removed by FCP1 phosphatase (Cho et al. 2001).

Reflecting its gradual enrichment at the gene body, S2p is essential for proper transcription elongation. For example, SETD2 methyltransferase which methylates histone H3 on Lys 36 (H3K36me3), binds to RNAPII CTD when S5p and S2p are present. H3K36me3 is enriched at the gene bodies of actively transcribed genes (Hampsey and Reinberg 2003) and was shown to supress intragenic transcription initiation in yeast (Carrozza et al. 2005; Venkatesh et al. 2012). Furthermore, S2p recruits SPT6, an elongation factor that coordinates nucleosome disassembly and reassembly during transcription and facilitates RNAPII elongation. S2p is also important for co-transcriptional recruitment of splicing factors: it was found associated with spliceosome components (Harlen et al. 2016) and mutating Ser2 residues to alanine impairs splicing due to defects in U2 auxiliary factor (U2AF) recruitment (Gu, Eick, and Bensaude 2013). Finally, S2p has an important role in 3'-end formation of mRNAs, which will be discussed in more detail below.

Another CTD modification that plays a role in elongation is threonine phosphorylation (T4p). It is enriched at the gene body and TES in both yeast and mammals, placed by PLK3 and CDK9 and removed by FCP1 phosphatase (Hintermair et al. 2012; Hsin, Sheth, and Manley 2011). In yeast, T4p appears to promote splicing, since phospho-specific immunoprecipitation of the T4p interactome is enriched for several elongation factors (Harlen et al. 2016), in chicken, it is linked with the 3' end processing of histone RNAs (Hsin, Sheth, and Manley 2011), and in human cells, it is associated with transcription elongation

and appears to regulate M-phase progression and chromosome segregation (Hintermair et al. 2012; Hintermair et al. 2016).

1.1.5 Termination

The last step of the transcription cycle is termination, where the nascent RNA is cleaved and polyadenylated and RNAPII dissociates from the DNA template shortly after the TES (reviewed in detail in (Porrua and Libri 2015; Proudfoot 2011)). Cleavage and polyadenylation of the nascent RNA depends firstly on the consensus poly(A) signal (PAS), AAUAA, and on auxiliary sequences flanking the PAS, and, secondly, on the co-transcriptional recruitment of cleavage and polyadenylation factors. Cleavage of the nascent RNA happens a few base pairs downstream of the PAS and is performed by cleavage and polyadenylation specificity factor (CPSF73), cleavage stimulatory factor (CstF) and cleavage factor 1 and 2 (CF1, CF2). Subsequently, poly(A) polymerase (PAP) adds a poly(A) tail to the 3'end of nascent RNA, which will not only protect the mRNA from degradation by exonucleases, but also favor its export to the cytoplasm and translation (Porrua and Libri 2015). Finally, RNAPII complexes dissociate from the DNA template and transcription terminates. There are two main models to explain RNAPII termination, but the detailed mechanisms are still unclear. The allosteric model proposes that termination occurs due to loss of elongation factors and/or conformational changes in elongating RNAPII (Proudfoot 2011; Zhang, Rigo, and Martinson 2015). The torpedo model postulates that termination is triggered by XRN2 exonuclease, which is recruited after cleavage of the nascent RNA, and degrades the nascent RNA up to the transcribing RNAPII, causing the dissociation of the elongating RNAPII complex (Loya and Reines 2016; Proudfoot 2011). Another important factor in transcription termination is RNAPII pausing downstream of the PAS at GC-rich sequences, where R-loops form. These are resolved by Senataxin helicase, which promotes XRN2 access to the nascent RNA and transcription termination (Skourti-Stathaki, Proudfoot, and Gromak 2011; Parua et al. 2018; Cortazar et al. 2019). Importantly, the allosteric and torpedo models are not mutually exclusive and a combination of both is likely

to happen *in vivo*. After transcription termination, RNAPII complexes are dephosphorylated and recruited to a new transcription cycle (Cho et al. 2001).

RNAPII CTD modifications also play a role in transcription termination. S2p interacts with and likely recruits cleavage and polyadenylation factors PCF11 and CPSF73 to the nascent RNA in yeast (Ahn, Kim, and Buratowski 2004; Gu, Eick, and Bensaude 2013; McCracken, Fong, Yankulov, et al. 1997; Meinhart and Cramer 2004). T4p was also recently shown to mediate transcription termination through the recruitment of RTT103 termination factor in yeast (Harlen et al. 2016). Finally, mutating the arginine residue (R1810) at the non-consensus repeat 31 of the CTD to alanine causes RNAPII accumulation and increased R-loop formation at the 3'end of genes, suggesting that it plays a role in termination in humans (Zhao et al. 2016).

1.2 Splicing

In eukaryotes, most genes are transcribed as a precursor mRNAs (pre-mRNAs), which are composed of coding sequences (exons) intercalated by stretches of non-coding sequences (introns). To produce an mRNA molecule, all introns must be removed and the exons joined together in a process called pre-mRNA splicing. Splicing depends on both *cis* and *trans* regulatory elements, which are regulated in a developmental-stage or tissue-specific manner. Proper splicing of pre-mRNA molecules is essential for appropriate gene expression and cellular function, and mis-spliced mRNAs are often linked with disease, such as neurodegenerative disorders and cancer (Matera and Wang 2014).

1.2.1 Canonical splicing

The pre-mRNA splicing reaction depends on several sequences present in all introns, which are highly conserved between yeast and mammals: GT, the donor site which defines 5' splice site (5'ss), AG, the acceptor site which defines 3' splice site (3'ss), and the branch point signal (BPS) which consists of 18-40 nucleotides upstream of the 3'ss. Metazoans contain an additional sequence that helps define

the 3'ss, called the polypyrimidine tract (PPT). The splicing reaction is a two-step transesterification reaction (**Fig. 1.3**).

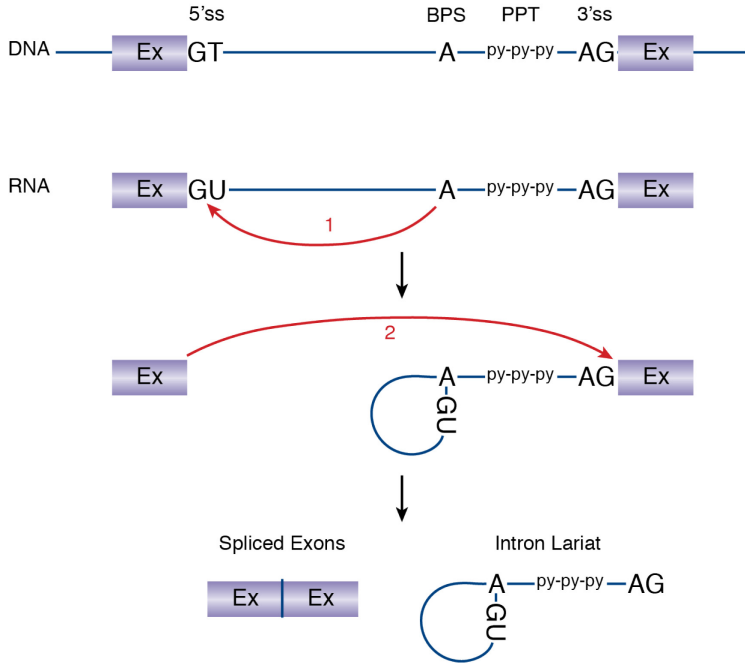

Figure 1.3 Illustration of the steps in the splicing reaction.
DNA sequence is depicted on top and RNA below. Red arrow 1 represents the first reaction step, where 2'OH of the adenosine performs a nucleophilic attack on the 5'ss. Red arrow 2 represents the second reaction step, where the free 3'OH of the 5'ss attacks the first nucleoside downstream of the 3'ss.

In the first step, the 2'OH of the adenosine in the BPS performs a nucleophilic attack on the 5'ss guanosine, producing a 5' exon with a free 3'OH and a branched intron lariat connected to the 3' exon. In the second step, the free 3'OH of the 5' exon attacks the first nucleotide downstream of 3'ss guanosine, yielding two spliced exons and an excised intron lariat (Herzel et al. 2017; Matera and Wang 2014; Will and Luhrmann 2011).

The splicing reaction is catalyzed by the spliceosome complex, which is composed of U-rich small nuclear ribonucleoproteins (snRNPs) named after their corresponding snRNA: U1, U2, U4, U5 and U6 (Matera and Wang 2014; Will and Luhrmann 2011). Most evidence supports a step-wise assembly of the spliceosome, where snRNPs dynamically assemble and disassemble throughout the different stages of splicing (**Fig. 1.4**).

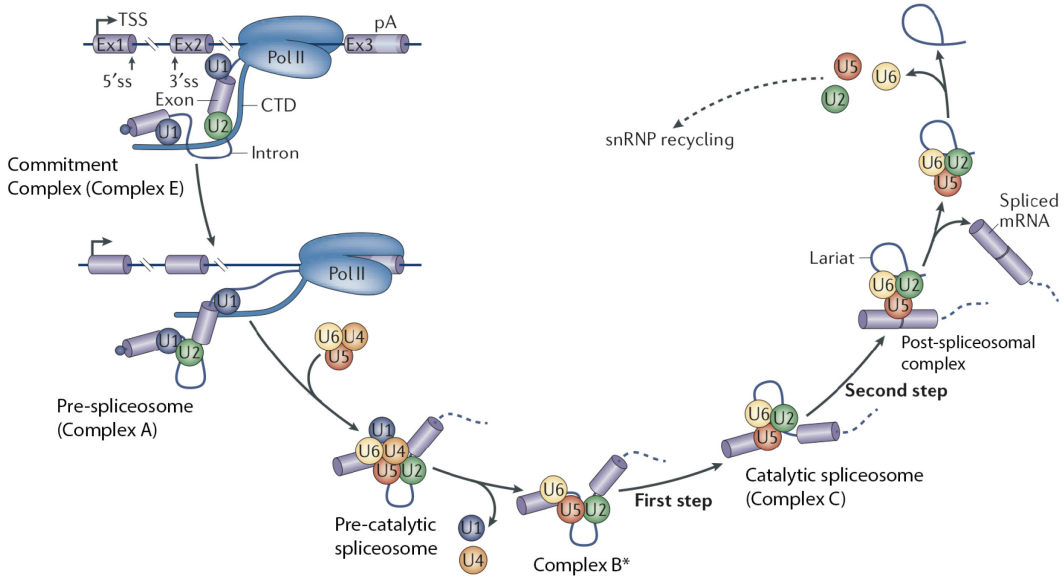

Figure 1.4 The splicing cycle.
Dynamic assembly and disassembly of snRNP complexes during the splicing cycle (Matera and Wang 2014).

Splicing itself does not require ATP, but ATPases are essential for the conformational transitions that occur during spliceosome assembly. Although this process is best studied in yeast, the main assembly steps are highly conserved in mammalian cells. Importantly, most splicing reactions are thought to occur co-transcriptionally, as the nascent RNA exits RNAPII. Splicing starts by the recognition of the 5'ss by the U1 snRNP, followed by Splicing Factor 1 (SF1) and U2AF binding to the BPS and the PPT at the end of the intron, thus forming the Commitment complex (complex E). Upon dissociation of SF1 and U2AF from the BPP and ATP hydrolysis, U2 snRNP recognizes the 3'ss and binds U1 snRNP, forming the Pre-spliceosome complex (complex A). After complex A assembly, the tri-snRNP, composed of U4, U5 and U6, is recruited to form the Pre-catalytic spliceosome, or complex B. Upon structural rearrangements catalyzed by several RNA helicases, U1 and U4 are evicted and the Activated complex B (complex B*) is formed. When the first catalytic step is completed, the Activated complex B gives rise to the Catalytic spliceosome, or complex C, which undergoes ATP-dependent rearrangements and performs the second step of the splicing reaction. Afterwards, the Post-spliceosomal complex is formed, which contains the spliced exons and an intron lariat. Finally, U2, U5, and U6 snRNPs are released from the

spliced exons, dissociate from the intron lariat and are recycled to a new splicing reaction.

1.2.2 Alternative splicing

Splicing is a highly complex process that is regulated by intricate networks that are cell-type and developmental stage specific. Constitutive exons are robustly recognized by the spliceosome and included in the mRNA. However, some exons may or may not be included in the mRNA and are alternatively spliced. Alternative splicing (AS) is the process by which different splice sites are recognized by the spliceosome. AS is important for numerous cellular processes, such as pluripotency, cell differentiation, circadian rhythm, response to environmental cues, or disease (Braunschweig et al. 2013). Approximately 95% of all transcripts were found to produce alternatively spliced mRNAs (Braunschweig et al. 2013) and complexity of AS events is related to increasing complexity of cell-types and species (Barbosa-Morais et al. 2012). Several types of AS have been described (Vuong, Black, and Zheng 2016), the most common being "cassette exons", where the exon is included or excluded from the mRNA (**Fig. 1.5A**). Exons can be mutually exclusive (**Fig. 1.5B**) or spliced from different 5'ss or 3'ss (**Fig. 1.5C and 1.5D**). mRNA isoforms can also have alternative last exons or poly(A) sites (**Fig. 1.5E and 1.5F**), and introns may be retained (**Fig. 1.5G**). Lastly, and a core topic of this thesis, the 5'ss of an exon can be spliced to an upstream 3'ss in a process called back-splicing, producing a circRNA molecule (**Fig. 1.5H**).

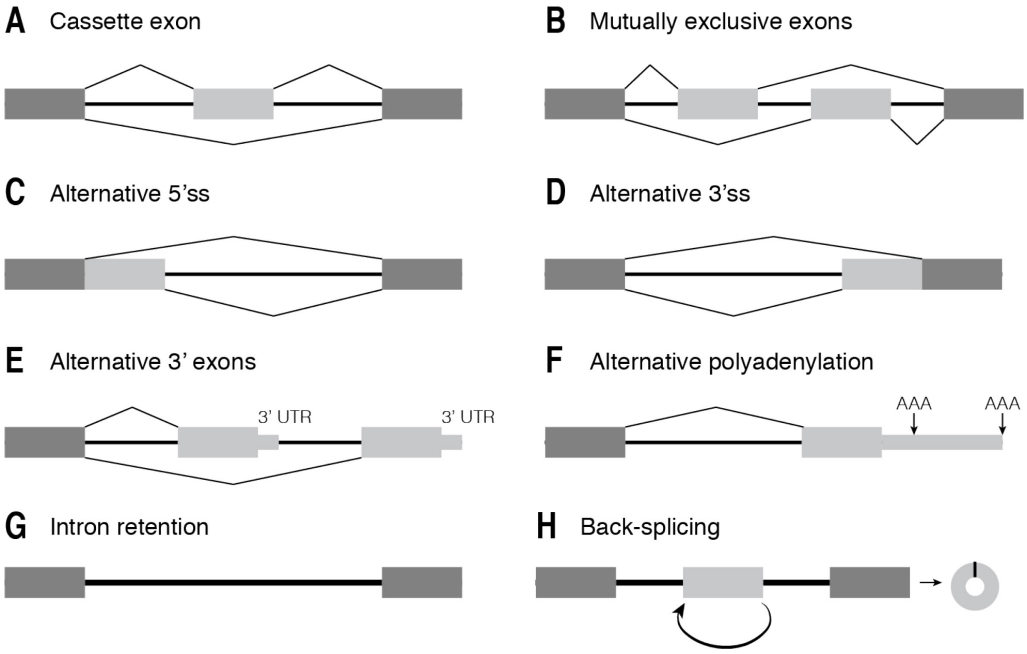

Figure 1.5 Types of alternative splicing.
RNA is depicted as a black line, constitutive exons are dark grey and alternatively spliced exons are light grey. **A)** "Cassette" exons which are included or excluded. **B)** Mutually exclusive exons. **C, D)** Alternative 5'ss or 3'ss selection. **E, F)** Alternative 3' exons or polyadenylation site usage. **G)** Intron retention. **H)** Back-splicing and circRNA formation (Vuong, Black, and Zheng 2016).

AS events are an important source of protein diversity, since AS can change protein domains and/or protein structure, thus impacting protein function. Selection of different last exons or PAS also influences which RNA binding proteins (RBPs) bind to the mRNA, thereby impacting mRNA stability, localization and translation.

Splicing is a highly regulated process which depends on numerous *cis* and *trans* factors, because canonical splicing sequences (GU/AG and BPS) are short and insufficient for appropriate splicing. For example, many introns contain cryptic splice sites, *i.e.* decoy sequences that resemble splice sites but are not functional, which can be recognized by the spliceosome and result in incorrect splicing (Braunschweig et al. 2013; Matera and Wang 2014). As such, splicing accuracy is increased by the presence of exonic and intronic splicing enhancer (E/ISE) and silencer (E/ISS) sequences in the vicinity of splice sites (Fig 1.6A). These sequences are bound by several RBPs, which promote or inhibit splice site recognition by the spliceosome.

Although splicing outcome firstly depends on the presence of *cis* elements, it is also highly dependent on the action of several RBPs binding to these elements. As several RBPs may bind to the same RNA molecule, their combined action can dictate a variety of splicing outcomes, that depends on the differential expression of RBPs (**Fig. 1.6B)** (Braunschweig et al. 2013; Vuong, Black, and Zheng 2016). There are two main classes of RBPs: constitutive and tissue-specific. Constitutive RBPs are expressed in almost all cells and tissues. Two large classes of constitutive RBPs are for example Serine/Arginine proteins (SR) and heterogeneous ribonucleoproteins (hnRNPs). Members of these protein classes bind to splicing enhancers or silencers and promote or inhibit splicing depending on their binding location and sequence context. Many other RBPs are tissue-specific and play an essential role in cellular function, such as PTBP2 in the brain or RBM24 in the heart. This is reviewed in detail in (Baralle and Giudice 2017).

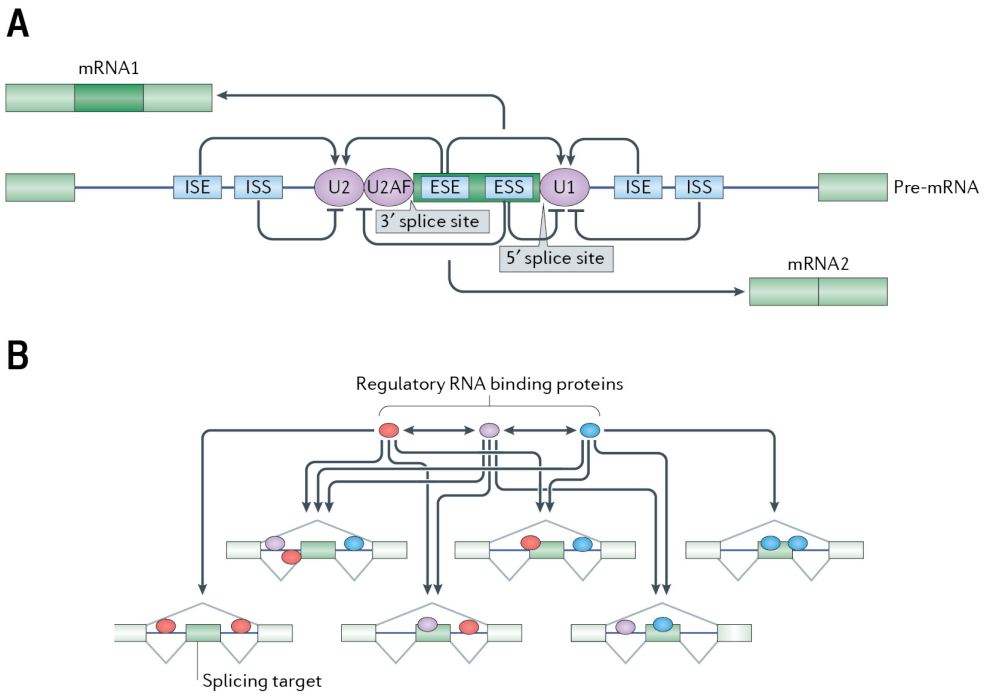

Figure 1.6 Splicing is regulated by cis and trans elements.
A) Schematics of the combined action of core spliceosome components and splicing enhancers and silencers that can influence splice site selection. **B)** Coloured circles represent different RBPs acting together to form highly complex splicing regulatory networks (Vuong, Black, and Zheng 2016).

Regulation of alternative splicing is particularly complex in neuronal tissues and several RBPs play an essential role in neurogenesis and neuronal function (reviewed in detail in (Vuong, Black, and Zheng 2016; Baralle and Giudice 2017)). An important and well-described example of RBP regulation is the interplay between polypyrimidine tract-binding protein 1 (PTBP1), PTBP2 and Serine/Arginine repetitive matrix protein 4 (SRRM4) during neurogenesis. PTBP1 is expressed in many cells, including neuronal stem cells and neuronal progenitor cells (NPCs), but not in neurons, where PTBP2 is expressed instead. In non-neuronal cells, PTBP1 inhibits the inclusion of exon 10 of PTBP2 mRNA, which forms a premature termination codon and triggers degradation of PTBP2 transcript by non-sense mediated decay. During neuronal maturation, when NPCs exit the cell cycle, PTBP1 is downregulated whereas SRRM4 is upregulated, thus promoting the inclusion of exon 10 of PTBP2 and its mRNA expression (Makeyev et al. 2007; Raj et al. 2014). In turn, PTBP2 plays an important role in neuronal development and maintenance. For example, it represses exon inclusion of proteins which regulate cell proliferation, cell fate and actin cytoskeleton (Licatalosi et al. 2012; Li et al. 2014). Not only does SRRM4 regulate PTBP2 mRNA expression, it also promotes the inclusion of neuronal microexons (Irimia et al. 2014), which regulate the function of several proteins during neurogenesis. This study also showed that microexons were found dysregulated in individuals with autism spectrum disorders, due to changes in SRRM4 expression.

1.2.3 Interplay between transcription and splicing

Another factor that greatly influences splicing outcome and is essential for appropriate splicing of nascent transcripts is the transcription machinery itself. Coupling splicing to transcription should allow the positioning of the spliceosome in close proximity to transcript splice sites as they exit RNAPII, as well as modulate competition between adjacent splice sites, thereby promoting optimal splice site selection and splicing reliability.

The transcription machinery facilitates splice site recognition

Recognition of splice sites and formation of the commitment complex occurs differently in short and long introns. In short introns (~250 bp), splice sites are thought to be recognized mainly through interactions between U1 and U2 snRNPs within the same intron and splicing commitment occurs via "intron definition" (Herzel et al. 2017; Hollander et al. 2016; Matera and Wang 2014). However, in long introns (~10 kb or larger), splice sites are much further apart and splice site recognition is thought to happen via "exon definition" (**Fig. 1.7**) (Berget 1995; Robberson, Cote, and Berget 1990; Schneider et al. 2010). This model postulates that U1 snRNPs and U2AFs would be associated with RNAPII CTD during transcription. As the nascent RNA exits RNAPII, the 5'ss would be recognized and bound by the first U1 snRNP complex, which would tether the nascent RNA to transcribing RNAPII until the synthesis of the 5'ss in the next intron. At this moment, a second U1 snRNP complex would be recruited to the nascent RNA via RNAPII CTD, recognize the 5'ss in the next intron and promote U2AFs binding to the 3' end of the previous intron, bringing the 5' and 3'ss of the same intron in close proximity and forming the Commitment complex. Finally, U2 snRNP would bind to the BPS, forming the Pre-spliceosome complex and the splicing reaction would occur.

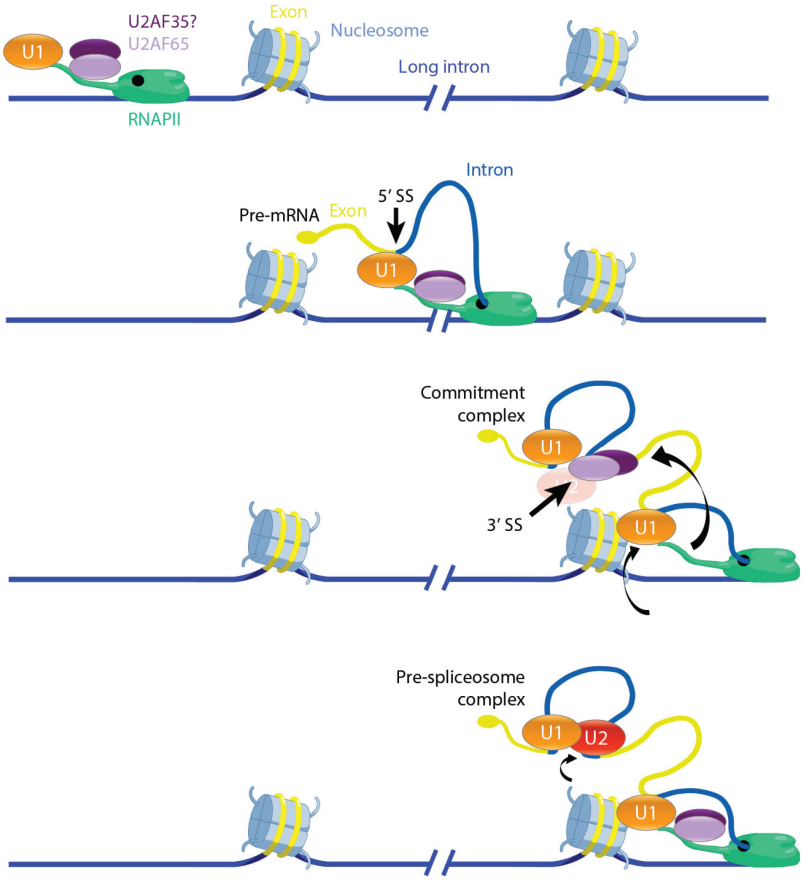

Figure 1.7 Illustration of the exon definition model.
DNA is depicted in dark blue, exons in yellow and introns in lighter blue (Hollander et al. 2016).

Intron definition is thought to be more prevalent in lower eukaryotes, where most introns are short, whereas exon definition is thought be more prevalent in higher eukaryotes, where most introns can be very large (Herzel et al. 2017; Matera and Wang 2014). Nevertheless, it is likely that both models occur depending on the intron length. For example, Pai and colleagues have compared the splicing dynamics of short and long introns in *Drosophila* S2 cells and showed that both shorter and longer introns are more efficiently spliced, whereas introns with intermediate length are spliced more slowly, suggesting that there is an optimal intron length for splicing via intron or exon definition(Pai et al. 2017). A recent study used fluorescently-tagged single pre-mRNAs to investigate the effect of multiple U1 snRNP complexes on splicing efficiency when forming cross-intron or cross-exon complexes (Braun et al. 2018). U2 binding to the 3'ss was found to be enhanced by the binding of U1 snRNP complexes to the 5'ss of the following

intron and that U1 snRNP complexes binding to both upstream and downstream 5'ss worked synergistically to promote U2 binding to the 3'ss recognition and subsequent splicing.

Several studies also support an intimate relationship between the establishment of splicing commitment and RNAPII regulation. RNAPII CTD was proposed to tether exons separated by long distances to promote accurate splicing (Dye, Gromak, and Proudfoot 2006). mNET-seq studies revealed that, in yeast, RNAPII was enriched at 5' and 3'ss (Mayer et al. 2015), whereas in mammals, RNAPII S5p was specifically enriched at 5'ss. This enrichment was splicing-dependent and these elongating RNAPII complexes were bound by U1 snRNP (Nojima et al. 2015; Nojima et al. 2018). Pulldowns of specific RNAPII modifications further confirm that RNAPII S5p is associated with spliceosome complexes (Harlen et al. 2016; Nojima et al. 2015; Nojima et al. 2018).

Recognition of the first exons requires additional mechanisms which also depend on the transcription machinery. The first exon is structurally distinct, as it contains a me7Gppp cap. The nuclear cap binding complex (CBC) binds to the capped nascent RNA and work as a recruitment platform for other factors that promote nascent RNA processing. For example, CBC directly interacts with U2-U5-U6 tri-snRNP and enhances its recruitment *in vivo*, and depletion of the CBC decreased U1 snRNP and U2-U5-U6 tri-snRNP recruitment to the nascent RNA (Gornemann et al. 2005; Lewis et al. 1996; Pabis et al. 2013). As previously discussed, RNAPII S5p is essential for nascent RNA capping and interacts with U1 snRNP; S5p is also most enriched at the TSS, when both capping and early spliceosome complexes are recruited to the nascent RNA, suggesting that RNAPII CTD plays an important role in facilitating the first splicing reaction. Finally, splicing of the last exon also depends on the presence of PAS and RNAPII. Removal of the PAS leads to readthrough transcription and inhibits splicing of the last exons and several spliceosome components interact with CPSF (Herzel et al. 2017), indicating that polyadenylation promotes splicing of the last exon by facilitating spliceosome assembly and vice versa. Furthermore, several RNAPII

CTD modifications, such as S2p and T4p, are essential for co-transcriptional recruitment of both splicing and cleavage and polyadenylation factors (Ahn, Kim, and Buratowski 2004; Gu, Eick, and Bensaude 2013; Harlen et al. 2016; Meinhart and Cramer 2004; McCracken, Fong, Yankulov, et al. 1997), which suggests that RNAPII CTD also mediates splicing of the last exon. Thus, spliceosome complexes interact with RNAPII CTD at every step of the transcription cycle and appropriate CTD modifications are essential for proper splice site recognition.

Crosstalk between transcription and splicing

Numerous studies have shown that transcription modulates alternative splicing, and two models were proposed to explain the results observed: the kinetic and the recruitment models. The "kinetic model" postulates that the elongation rate of RNAPII complexes regulates alternative splicing by providing more or less time for the splicing machinery to recognize different splice sites (Braunschweig et al. 2013; Hollander et al. 2016; Kornblihtt et al. 2013). For example, fast elongating RNAPII complexes would not provide enough time for the spliceosome to recognize a weak splice site, thereby promoting exon exclusion (**Fig. 1.8A)**. Conversely, slow elongating RNAPII complexes would allow more time for the spliceosome to recognize a weak splice site and promote exon inclusion (**Fig. 1.8B**). However, this is not always the case, since the effects of RNAPII elongation rate on splicing greatly depend on which splicing regulators have the opportunity to bind to the nascent RNA. For example, (Dujardin et al. 2014) have shown that slow elongating RNAPII facilitates the recruitment of the ETR-3 splicing inhibitor, promoting exclusion of exon 9 in CFTR mRNA. Extensive evidence supports the kinetic model. Exons have distinct chromatin features when compared to introns, such as increased nucleosome density, enrichment for specific histone modifications, or increased CpG methylation, which negatively correlate with RNAPII elongation rate in mammalian cells and are thought to help define exons (Jonkers, Kwak, and Lis 2014; Veloso et al. 2014).

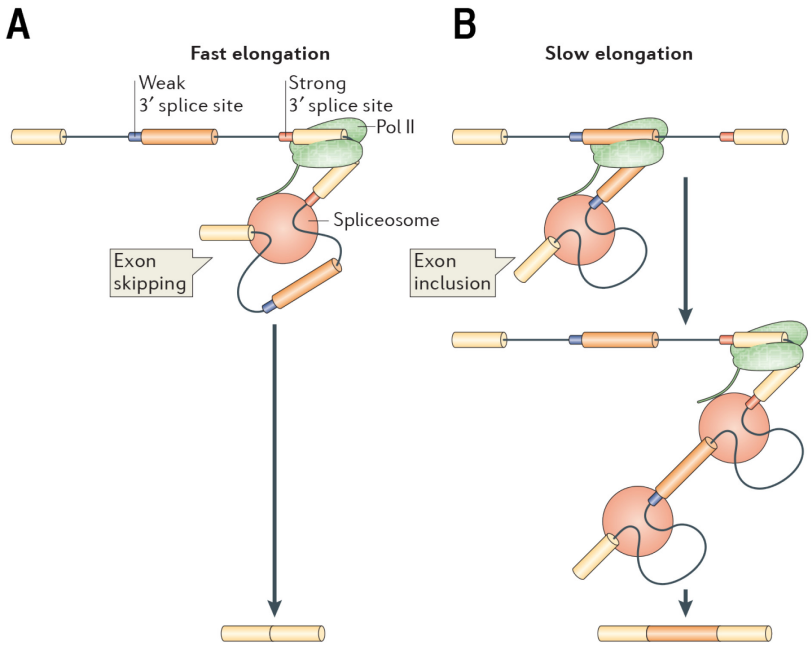

Figure 1.8. Kinetic model.
Here are depicted the effects that **A)** fast or **B)** slow elongation rate of RNAPII on alternative splicing (Kornblihtt et al. 2013).

Modulation of RNAPII elongation rate (by either drug treatments such as DRB or campthothecin, DNA damage, or slow/fast RNAPII elongation mutants) revealed extensive changes in splicing patterns (de la Mata et al. 2003; Ip et al. 2011; Maslon et al. 2019; Munoz et al. 2009). Importantly, evidence suggests that there is an optimal RNAPII elongation rate is thought to exist for appropriate splicing outcome (Fong et al. 2014; Aslanzadeh et al. 2018). Finally, factors that cause stalling of RNAPII near exons tend to promote exon inclusion (Shukla et al. 2011). For example, CTCF binds to unmethylated CpG islands, thereby stalling RNAPII and promoting exon inclusion.

The "recruitment model" postulates that promoter architecture and/or the RNAPII CTD can recruit splicing factors which influence splice site selection (Braunschweig et al. 2013; Hollander et al. 2016; Kornblihtt et al. 2013). Several studies showed that the type of promoter (Cramer et al. 1997; Cramer et al. 1999), transcriptional activators (Kadener et al. 2001; Nogues et al. 2002) and chromatin remodelers (Batsche, Yaniv, and Muchardt 2006) can impact alternative splicing. For example, the mediator complex regulates the recruitment of hnRNP L, which in turn causes exon exclusion (**Fig. 1.9A)** (Huang et al. 2012).

Finally, as previously discussed, RNAPII CTD is essential for proper splicing (McCracken, Fong, Yankulov, et al. 1997; Ryan et al. 2002; Rosonina and Blencowe 2004) and recruits both core spliceosome components and alternative splicing factors, such as SR proteins, to the nascent RNA (**Fig. 1.9B)**, (Das et al. 2007; David et al. 2011; Morris and Greenleaf 2000; de la Mata and Kornblihtt 2006; Harlen et al. 2016; Nojima et al. 2018). Importantly, the kinetic and recruitment models are not mutually exclusive and a combination of both is likely to occur *in vivo*.

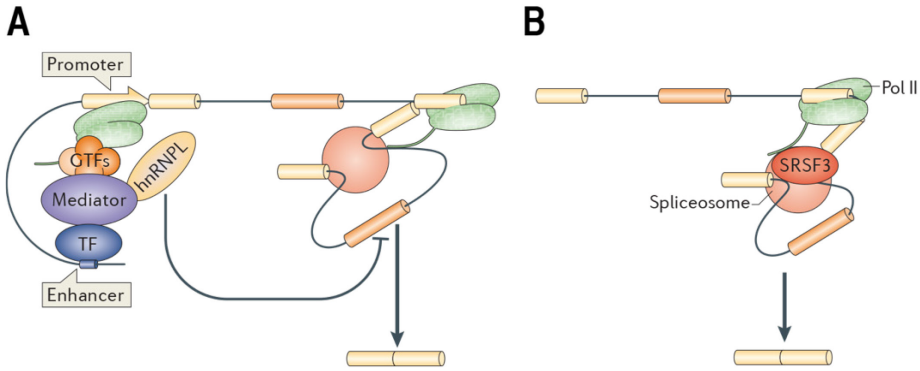

Figure 1.9 – Recruitment model.
Alternative splicing regulators can be recruited by **A)** promoters and **B)** the CTD of RNAPII (Kornblihtt et al. 2013).

The relationship between transcription and splicing does not seem to be unidirectional. In fact, several studies hint at splicing kinetics reaching back to modulate transcription. For example, splicing was shown to cause RNAPII stalling at splice sites (Mayer et al. 2015; Nojima et al. 2015; Milligan et al. 2017) and introns (Alexander et al. 2010; Chathoth et al. 2014). Furthermore, splicing inhibition affects transcription. For instance, inhibiting splicing with spliceostatin A or pladienolide B decreased RNAPII S2p cellular levels (Koga, Hayashi, and Kaida 2015) and knockdown of SC35 splicing factor increased RNAPII density at the gene body with decreased S2p at *Ptbp1* and *Ets1* genes (Lin et al. 2008). Finally, a recent paper by Akhtar and colleagues in *Drosophila* has linked splicing to promoter-proximal pausing regulation (Akhtar et al. 2019). Depleting subunits of the pre-Exon Junction Complex (pre-EJC), which binds introns and aids in intron definition, caused decreased total RNAPII at promoters with increased S2p

at the TES together with splicing defects. Genes with high levels of paused RNAPII at promoters seemed to be more affected by the knockdown of pre-EJC. The authors further showed that the pre-EJC directly interacts with RNAPII S5p in an RNA-dependent manner and that tethering of pre-EJC subunit Mago to the 5' end of nascent RNAs is sufficient to rescue promoter-proximal pausing of RNAPII and splicing defects. Some of these observations were extended to human cells. It was proposed that the pre-EJC stabilizes RNAPII pausing at the promoter by restricting PTEF-b access to RNAPII, thereby providing enough time for the recruitment of additional splicing factors essential for proper exon definition.

Extent of co-transcriptional and post-transcriptional splicing

Although splicing factors are thought to be co-transcriptionally recruited, the splicing reaction is not necessarily completed co-transcriptionally. Indeed, the extent to which splicing is co-transcriptional is a highly debated matter. Several studies based on RNA-seq (bulk, single-molecule, long-read nascent RNA) and mNET-seq indicate that most splicing occurs co-transcriptionally (Ameur et al. 2011; Herzel, Straube, and Neugebauer 2018; Khodor et al. 2012; Nojima et al. 2018; Oesterreich et al. 2016; Tilgner et al. 2012). For example, sequencing of RNAs protected by RNAPII together with spliceosome complexes in mammalian cells revealed that a large proportion of RNAs were spliced, suggesting that most RNAs are co-transcriptionally spliced (Nojima et al. 2018). However, other studies quantifying the rate of splicing in living cells with orthogonal techniques, such as qPCR (Singh and Padgett 2009), single-molecule imaging (Coulon et al. 2014; Martin et al. 2013), or metabolic labeling of RNA (Rabani et al. 2014; Windhager et al. 2012) estimate that splicing reactions may occur between 2 min to 1h after transcription, hinting that post-transcriptional splicing is possible to a certain extent.

The assumption that splicing mostly happens co-transcriptionally also implies that the order of splicing should follow the "first come, first served" model, which postulates that introns would be spliced as RNAPII transcribes the gene. In yeast, splicing appears to occur very rapidly and as soon as introns exit the RNAPII

channel (Herzel, Straube, and Neugebauer 2018; Oesterreich et al. 2016). However, in more complex organisms, this does not appear to be as straightforward. For example, a study by (Khodor et al. 2012) showed that the closer introns are to the 3' end of the gene, the lower is the likelihood of co-transcriptional splicing. Vargas and colleagues suggested that although constitutive exons were mostly co-transcriptionally spliced, alternatively spliced exons were post-transcriptionally spliced and that reducing splicing efficiency caused introns to be spliced post-transcriptionally (Vargas et al. 2011). Finally, a preprint from the Churchman lab (Drexler, Choquet, and Churchman 2019) has addressed this issue systematically in *D. melanogaster* and human cells with Co-Transcriptional Processing, a new method that determines long nascent RNA isoforms without amplification biases. They report that *D. melanogaster* introns tend to be co-transcriptionally spliced and in the order of the transcript. On the contrary, in human K562 cells, splicing seems to occur after RNAPII has transcribed several kb, suggesting an increased tendency for post-transcriptional splicing. Further, splicing frequently did not occur in the order of the linear transcript, and similar patterns were observed for the same mRNA isoforms, suggesting that these events were regulated. This study also indicated that cleavage and polyadenylation most often happened before the last splicing reaction was completed, in both *D. melanogaster* and human cells.

1.3 Circular RNAs

CircRNAs were initially discovered in in plants as viroid RNA particles (Sanger et al. 1976) and have have recently received increased interest with their discovery through genome-wide sequencing methods (Salzman 2016).CircRNAs are formed when exons are not linearly spliced but covalently linked by a 3'-5' phosphodiester bond to an upstream intron-exon junction of the same RNA molecule, in a process called back-splicing (**Fig. 1.10)**. CircRNAs are not polyadenylated and are resistant to degradation by RNAse R, an enzyme which degrades linear RNAs (Salzman 2016).

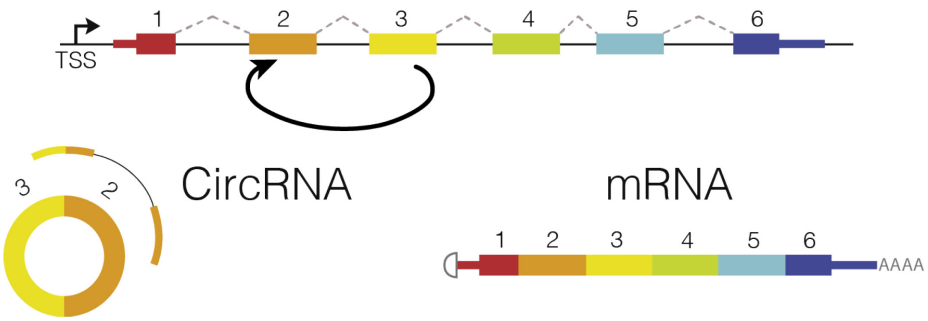

Figure 1.10 Schematic representation of circRNA formation through back-splicing.
DNA and intronic sequences are represented by the black line. Exons are depicted as coloured boxes. Dashed lines represent linear splicing and thick curved arrow represents back-splicing. Bottom left diagram represents a circRNA produced from exons 2 and 3 and the smaller curved lines represent back-spliced reads found in total RNA-seq. Bottom right diagram represents the corresponding linear transcript or mRNA (Salzman 2016).

The first circRNA identified was a viroid RNA particle in plants (Sanger et al. 1976) and, in the 1990s, other sporadic studies reported RNA molecules with "scrambled exons" in higher eukaryotes (Capel et al. 1993; Cocquerelle et al. 1992; Nigro et al. 1991). These were thought to be rare events and/or by-products of noisy splicing. The idea that most splicing happened co-transcriptionally rendered the concept of back-splicing highly unlikely and further contributed to dismissing these observations. However, the onset of next-generation sequencing technologies which recovered non-polyadenylated RNAs together with the development of computational algorithms that detected exons in a "scrambled" order revealed that circRNAs were a much more widespread phenomenon than originally thought (Jeck et al. 2013; Memczak et al. 2013; Salzman et al. 2012). Indeed, circRNAs are found in most eukaryotes, such as metazoans, fungi, protists and plants, some being conserved between species (Li, Yang, and Chen 2018; Wilusz 2018). CircRNAs are typically produced from protein-coding genes and contain 1-5 exons, with the internal introns removed and are often produced from the 5' end of genes (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012). CircRNAs are mostly cytoplasmic, although some remain in the nucleus (Salzman et al. 2012; Jeck et al. 2013; Li et al. 2015). CircRNAs are thought to be much more stable than linear mRNAs since they lack 5' and 3' ends, thereby having increased resistance to the regular mRNA decay pathways. Enuka and colleagues confirmed this by estimating that the average

half-life of circRNAs in mammalian cells ranged from 18 to 23h, compared with 4 to 7.4h of linear transcripts (Enuka et al. 2016). Nevertheless, the steady state levels of circRNAs positively correlate with the synthesis of nascent circRNAs, *i.e.* the more circRNAs are produced the higher the steady state levels of circRNAs detected (Conn et al. 2015; Salzman et al. 2013; Starke et al. 2015; Zhang et al. 2016). Additionally, circRNAs are mostly expressed in a cell-type specific manner and their expression levels are overall low when compared with their corresponding linear transcripts. Nevertheless, some circRNAs are more abundant than their corresponding linear transcripts and appear to negatively correlate with linear transcript expression (Jeck et al. 2013; Guo et al. 2014; Maass et al. 2017; Rybak-Wolf et al. 2015; Zhang et al. 2016). However, these analyses compare the expression of individual circRNAs with the expression of the corresponding host transcript and the relationship between circRNA expression with mRNA expression from a given gene has not been extensively explored.

CircRNAs are found in next generation sequencing datasets through computational algorithms that search for "back-spliced" reads, *i.e.* spliced reads that do not match the linear transcriptome. Most often this search is performed in total RNA-seq datasets depleted for ribosomal RNA that were very deeply sequenced (> 200 million reads). Several laboratories developed pipelines to identify circRNAs, such as find_circ (Memczak et al. 2013), CIRCexplorer (Zhang et al. 2014), circRNAFinder (Westholm et al. 2014) or CIRI (Gao, Wang, and Zhao 2015). Despite computational methods being critical for circRNA identification, false positives may occur and it is important to validate these events experimentally. Several methods can be used to validate predicted circRNAs (reviewed in detail in (Li, Yang, and Chen 2018)). One method consists of designing divergent primers which amplify the back-spliced junction of the circRNA of interest followed by qPCR quantification. Another method to detect circRNAs directly in the RNA population is by northern blot, using probes that specifically hybridize to the back-spliced junction or to both linear and circRNA

isoforms. Both northern blot and qPCR can be coupled with pre-treating the RNA sample with RNAse R, which degrades linear RNAs without affecting circRNA levels.

1.3.1 Function of circular RNAs

The biological roles of circRNAs are currently under intensive research. The best described role of circRNAs is as microRNA (miRNA) sponges. CDR1as is a single exon circRNA, which is highly conserved and abundant in the mammalian brain (Hansen et al. 2011; Hansen et al. 2013; Memczak et al. 2013). CDR1as is covered of mir-7 binding sites and its knockdown was shown to decrease the expression levels of mir-7 target mRNAs, suggesting that CDR1as competes with target mRNAs for mir-7 binding (Hansen et al. 2013; Memczak et al. 2013). The circRNA produced from *Sry* gene, which determines sex in mammals, has mir-138 binding sites and interacts with mir-138 in murine cells (Hansen et al. 2013). Nevertheless, most circRNAs are not enriched for miRNA binding sites, suggesting that decoy of miRNAs is not a common function (Enuka et al. 2016; Guo et al. 2014).

The formation of circRNAs has also been discussed in the context of modulating the formation of linear mRNA species. Some studies hinted that the usage of 5'ss and 3'ss in circRNA formation may compete with linear splicing, resulting in lower levels of the corresponding mRNA transcripts (Ashwal-Fluss et al. 2014; Zhang et al. 2014). Likewise, the ratio between the number of back-spliced and linearly spliced reads, a measure of circularization levels at specific exon junctions, negatively correlates with gene expression, suggesting competition between back-splicing and linear splicing at specific splice junctions (Rybak-Wolf et al. 2015). Back-splicing was also proposed to be associated with exon skipping events, where the skipped exon/s would produce circRNAs. Using primary endothelial human cells, Kelly and colleagues suggested that circRNA formation is correlated with exon skipping (Kelly et al. 2015). A different study using cardiac cells found that circRNAs appeared to mostly originate from constitutive exons

(Aufiero et al. 2018). As such, further studies are necessary to clarify whether a particular relationship exists between circRNA formation and exon-skipping in mRNA species.

More recently, the coding potential of circRNAs has also been investigated, and found that they can be translated. CircZNF609 was found to be translated and to play a role in myoblast differentiation of mammalian cells (Legnini et al. 2017). Ribosome profiling from fly heads also showed that a subset of circRNAs was associated with translating ribosomes and that circMbl produced a protein (Pamudurti et al. 2017). Yang and colleagues further showed that circRNAs are enriched for N^6-methyladenosine residues and that these residues may promote translation (Yang et al. 2017).

There are two additional subtypes of circRNAs, produced from alternative pathways, which appear to modulate gene expression of their host genes. Zhang and colleagues have described a type of circRNAs that originate from long introns which were not debranched (Zhang et al. 2013). An abundant species of these circRNAs, ci-ankrd52, accumulates at its transcription site and modulates the expression of its host gene by associating with RNAPII and promoting elongation. Another subclass of circRNAs is produced from back-splicing events with retained introns (exon-intron-circRNAs or EIciRNAs) (Li et al. 2015). EIciRNAs accumulate at the site of transcription and were shown to promote the expression of their host genes by binding to U1 snRNP and RNAPII at the TSS, thereby stimulating transcription and splicing.

CircRNAs have also been found to have roles in the innate immune response and cancer (reviewed in detail in (Chen, Satpathy, and Chang 2017) and (Patop and Kadener 2018)). For example, in human cells, immune-response factor NF90 (or its isoform NF110) promotes circRNA formation. Upon viral infection, NF90/NF110 shuttles to the cytoplasm to bind viral transcripts and suppress their transcription, causing decreased circRNA formation (Li et al. 2017). CircRNA production is also increased during human epithelial to mesenchymal

transition (Conn et al. 2015) and in prostate cancer (Chen et al. 2019). Importantly, circRNA levels appear to positively correlate with prostate cancer progression and a subset of circRNAs promote cell proliferation (Chen et al. 2019). Given their potential role in disease, circRNAs are emerging as promising biomarkers for disease diagnosis and prognosis (Patop and Kadener 2018).

Finally, circRNAs are increasingly found to play a role in neuronal function. They are most abundant in neuronal cells and tissues (Rybak-Wolf et al. 2015; You et al. 2015; Zhang et al. 2016). CircRNAs are also upregulated during neurogenesis, during *in vitro* neuronal differentiation, and in aging neuronal tissues of flies and mice (Gruner et al. 2016; Westholm et al. 2014; Rybak-Wolf et al. 2015; Zhang et al. 2016). Some circRNAs are much more abundant than their corresponding linear transcript and tend to accumulate in neuronal cells. This could be explained by both increased circRNA production from neuronal specific genes and high resistance of circRNAs to degradation mechanisms (Rybak-Wolf et al. 2015; You et al. 2015).

Further, accumulation of circRNAs was proposed to be more prominent in slow- or non-dividing cells (Rybak-Wolf et al. 2015; Zhang et al. 2016). A positive correlation was found between the number of circRNAs detected in a given cell type and its corresponding cellular division rate, *i.e.* fewer circRNAs are detected in cells with a faster division rate, compared to slowly dividing cells. This effect is thought to result from the dilution of circRNA molecules in a population of fast-dividing cells (Bachmayr-Heyda et al. 2015; Patop and Kadener 2018; Zhang et al. 2016). Some circRNAs were also found enriched in synaptosomes, after biochemical purification enriched in synapses, and their expression can be modulated by synaptic activity (Rybak-Wolf et al. 2015; You et al. 2015), suggesting that some circRNAs play a role in synaptic function. Indeed, CDR1as-knockout mice showed defects in excitatory synaptic plasticity and behavioral changes associated with neuropsychiatric disorders (Piwecka et al. 2017). Finally, a group of circRNAs in mouse stem cell-derived spinal motor neurons was shown to be regulated by FUS protein in mouse stem cell-derived spinal motor

neurons, an RBP implicated in Amyotrophic Lateral Sclerosis (ALS) (Errichelli et al. 2017) and some of these circRNAs were found deregulated in ALS patients with mutations linked to the FUS gene.

1.3.2 Biogenesis of circular RNAs

Circular RNA biogenesis is a highly complex process which is thought to be a rare by-product of linear splicing mechanisms and remains poorly understood. CircRNA formation is a highly unfavored event, where a U1 snRNP complex bound to a downstream 5'ss finds a U2 snRNP complex bound to an upstream 3'ss, followed by recruitment of the tri-snRNP complex and back-splicing (**Fig. 1.11A**). Amongst other complexities, the back-splicing reaction requires that the intron upstream to the first back-spliced exon is not removed before the synthesis of the splice donor downstream to the last exon included in the circRNA.

Several studies have found that both *cis* and *trans* factors contribute to and can modulate circRNA formation. Firstly, circRNA formation depends on the presence of canonical splice site sequences and their mutation impairs back-splicing (Ashwal-Fluss et al. 2014; Starke et al. 2015). Nevertheless, how exactly snRNPs assemble to promote back-splicing is still unknown. Another common feature of circRNAs, from *C. elegans* to humans, is that circularized exons tend to be flanked by very long introns which often also contain complementary sequences that anneal to each other (**Fig. 1.11B**) (Salzman et al. 2012; Ashwal-Fluss et al. 2014; Ivanov et al. 2015; Zhang et al. 2014). This RNA pairing is thought to bring the downstream 5'ss and the upstream 3'ss in spatial proximity and facilitate back-splicing, which often occurs at repetitive elements such as Alu repeats in primates (Jeck et al. 2013; Zhang et al. 2014) but can also originate from non-repetitive complementary sequences (Ashwal-Fluss et al. 2014; Ivanov et al. 2015; Liang and Wilusz 2014). Pairing of complementary sequences can be predictive of back-splicing events and 30-40 nucleotides appear sufficient to promote back-splicing in expression vectors (Zhang et al. 2014; Ivanov et al. 2015). Disruption of RNA

pairing by deletion of one of the sequences within the pair reduces back-splicing efficiency (Liang and Wilusz 2014; Zhang et al. 2014). If several complementary sequences are present, competition between distinct pairs can give rise to different circRNA isoforms in a process called alternative circularization (**Fig. 1.11C**) (Zhang et al. 2014). In lower vertebrates, however, the presence of complementary sequences seems to contribute far less to circRNA production because most introns flanking circularized exons are not enriched for complementary sequences, despite still being very long (Westholm et al. 2014). Although intronic pairing is important for circRNA formation, not all back-splicing events are explained by complementary sequences. Furthermore, circRNA expression differs considerably between cell types and tissues, suggesting that back-splicing does not depend on sequence alone and that *trans* factors regulate circRNA formation.

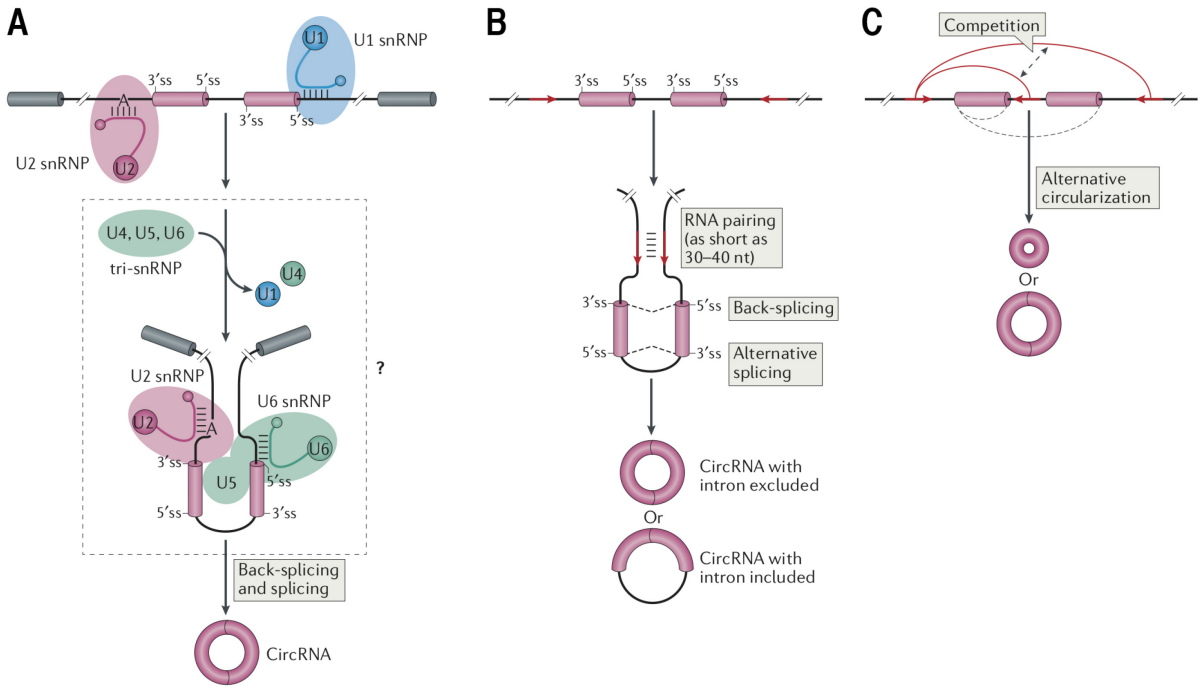

Figure 1.11 Regulation of circRNA formation by *cis* factors.
Introns are depicted as a black line, circularized exons are coloured in pink and other exons in grey. Complementary sequences are depicted as red arrows and RNA pairing as short black lines. **A)** CircRNA formation depends on canonical splicing sequences and spliceosome function. How the several snRNPs assemble to promote back-splicing is still unknown (highlighted by the dashed-line box). **B)** RNA pairing from intronic complementary sequences promotes circRNA formation. **C)** Competition between distinct RNA pairs can cause alternative circularization (Chen 2016).

Several studies have shown that circRNA formation can be modulated by RBPs, which may promote or disrupt intronic pairing preceding the back-splicing reaction. The first RBP shown to modulate circRNA formation was Muscleblind (MBL) in *D. melanogaster*: MBL binds to intronic target sequences in its own pre-mRNA and promotes circMbl production, thereby fine-tuning its own gene expression (**Fig. 1.12A**) (Ashwal-Fluss et al. 2014). Following this study, many RBPs were shown to favor circRNA formation through a similar mechanism in various biological contexts. Upon epithelial to mesenchymal transition in HMLE cells, Quaking (QKI) protein is upregulated and binds to single stranded RNA motifs flanking circularized exons. Dimerization of QKI is thought to promote intronic RNA pairing and stimulate circRNA formation (**Fig. 1.12B)** (Conn et al. 2015). Furthermore, crosslinking immunoprecipitation (CLIP) assays in mESC-derived spinal motor neurons found the FUS protein enriched at introns flanking circularized exons, thus promoting circRNA production (Errichelli et al. 2017). In HeLa cells, the immune-response factor NF90/NF110 was shown to bind to double-stranded RNA (dsRNA) and favor back-splicing (Li et al. 2017). Other RNPs which are general splicing regulators, such as hnRNPs and SR proteins, were also shown to regulate circRNA production (Fei et al. 2017; Kramer et al. 2015). On the contrary, some RBPs seem to prevent circRNA formation by destabilizing RNA pairing. For instance, the Adenosine deaminase acting on RNA 1 (ADAR1) enzyme, which converts adenosines into adenines (A-to-I editing), inhibited circRNA formation by disrupting intronic RNA pairs, and its knockdown increased circRNA production (**Fig. 1.12C)** (Ivanov et al. 2015; Rybak-Wolf et al. 2015). RNA helicase DexH-Box Helicase 9(DHX9) was also shown to inhibit production of circRNAs flanked by Alu repeats possibly through direct binding to dsRNA and unwinding RNA pairing, or through the recruitment of ADAR enzymes to these locations (Aktas et al. 2017).

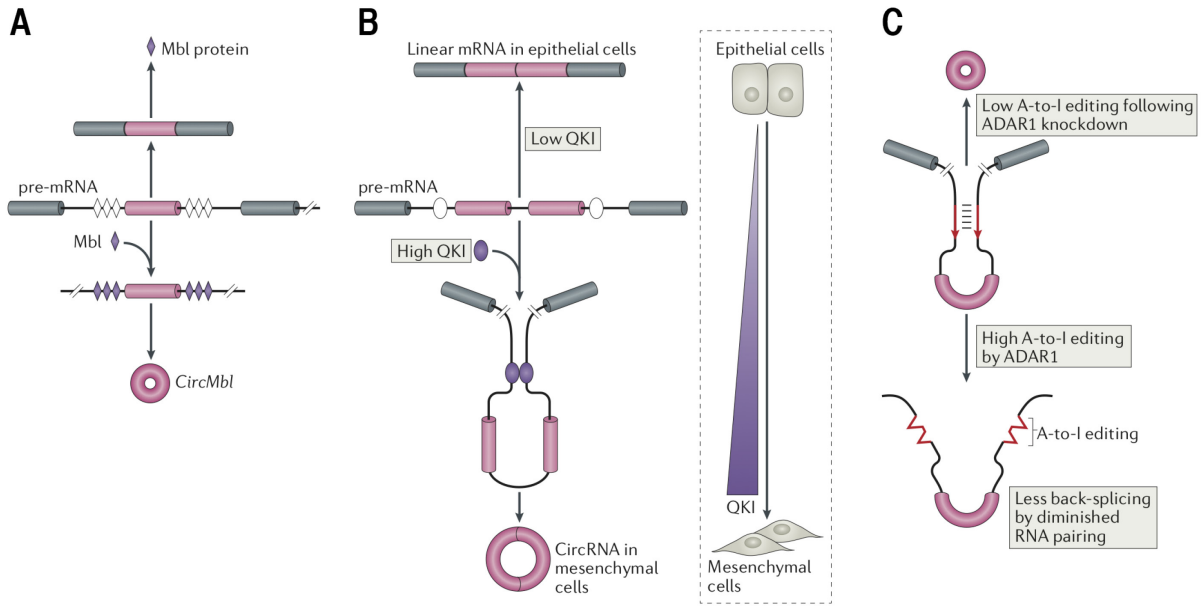

Figure 1.12 CircRNA regulation by *trans* factors.
Introns are depicted as a black line, circularized exons are coloured in pink and other exons in grey. Complementary sequences are depicted as red arrows and RNA pairing as short black lines. RBPs are represented in purple shapes and RBP binding sites in matched white shapes. A) MBL regulates the expression of its own host gene through promoting circMbl formation. B) Linear mRNAs are produced in epithelial cells in the presence of low QKI expression levels. In mesenchymal cells, where QKI is upregulated, QKI stimulates circRNA production. C) ADAR1 performs A-to-I editing and disrupts intronic RNA pairing, thus preventing circRNA formation (Chen 2016).

Finally, evidence suggests that circRNA production may be intimately linked with co-transcriptional splicing regulation. The first hint arose from Ashwal-Fluss and colleagues, where some circRNAs were found associated with chromatin-bound RNA from fly heads and that flies with a slow elongating RNAPII mutant, which was shown to increase splicing fidelity, had fewer circRNAs compared to wild type (Ashwal-Fluss et al. 2014). These findings were later confirmed in human cells (Zhang et al. 2016), where it was shown that nascent and steady state circRNAs often originated from genes with a fast elongating RNAPII (**Fig. 1.13A**). Zhang and colleagues further used fast and slow RNAPII mutants with α-amanitin (which inhibits transcription from wild type RNAPII) to dissect the relationship between elongation speed of RNAPII and circRNA production (Zhang et al. 2016). CircRNA levels increased in the presence of a fast elongating RNAPII, while the

opposite was observed for the slow RNAPII mutant. The authors suggested that fast RNAPII elongation might allow pairing of complementary sequences across introns instead of within introns, thereby facilitating back-splicing. Another recent study explored the role of transcription and splicing fidelity in circRNA formation in *D. melanogaster* cells by analyzing the knockdown effects of cleavage and polyadenylation factors and spliceosome components on circRNA production from reporter constructs and some endogenous loci (Liang et al. 2017). Results showed that knockdown of cleavage and polyadenylation factor Cpsf73 impaired cleavage and polyadenylation and transcription termination. This resulted in RNAPII continued transcription into the neighboring gene, a process called "*read-through transcription*", which promoted circRNA formation from the neighboring gene (**Fig. 1.13B**) (Liang et al. 2017). Knockdown of several spliceosome components, such as the U1C subunit of U1 snRNP, which also increased circRNA formation (**Fig. 1.13C**), suggesting that reduced splicing fidelity switched transcription output from linear RNA to circRNA production (Liang et al. 2017). This effect was observed for circRNAs containing only one exon. In summary, the above-mentioned studies indicate that the interplay between transcription and splicing is an important factor underlying circRNA formation.

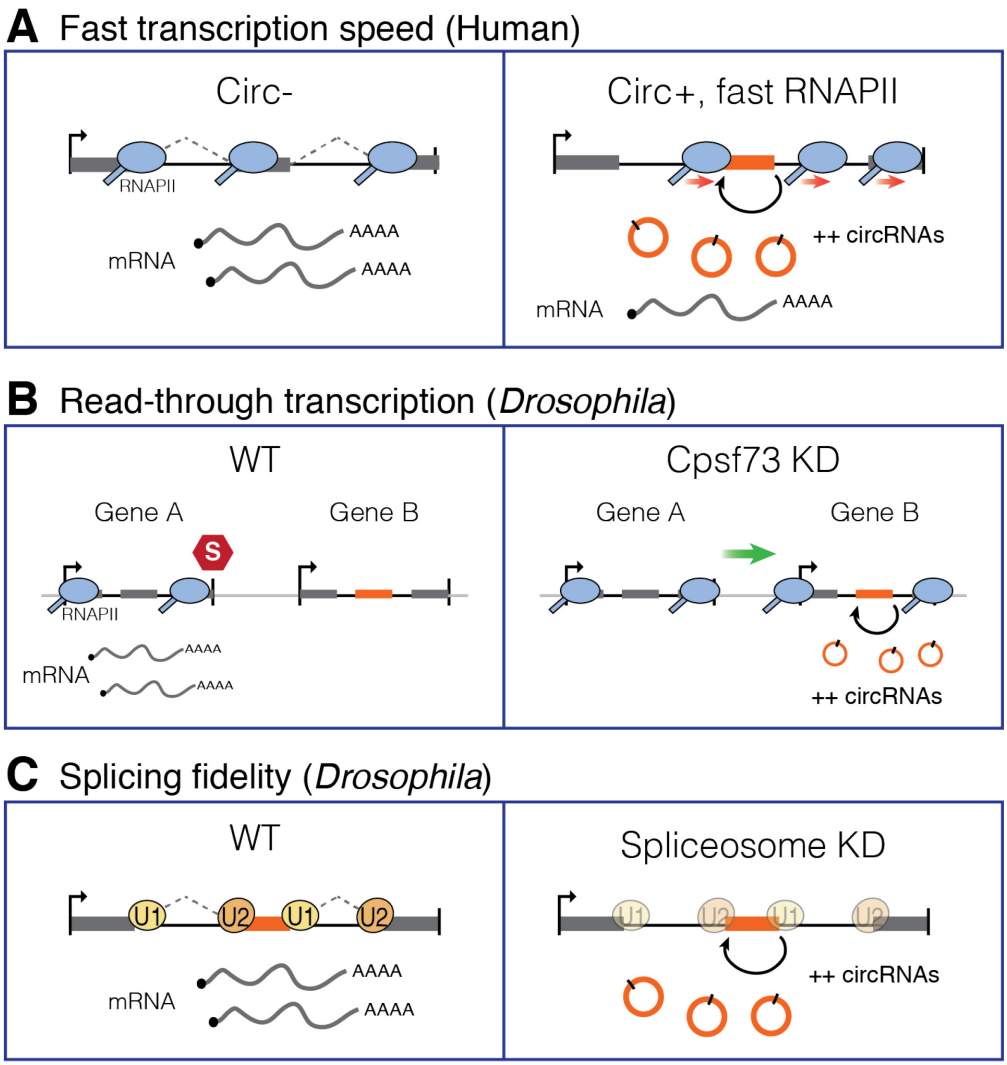

Figure 1.13 Transcription and splicing dynamics play a role in circRNA formation.
DNA and introns are depicted as a black line. Orange boxes are back-spliced exons and grey boxes are other exons. Dashed grey lines represent linear splicing and curved arrow back-splicing. RNAPII complexes are in blue and circRNAs are orange circles. **A)** CircRNA-producing (circ+) genes have fast-transcribing RNAPII complexes (red arrows), compared to genes not producing circRNAs (circ-). **B)** In wildtype conditions, RNAPII transcribes gene A and cleavage, polyadenylation and termination occur. Knockdown of Cpsf73 causes read-through transcription (green arrow) from gene A to B and triggers circRNA formation in *Drosophila*. **C)** In limiting spliceosome conditions, e.g. knockdown of spliceosome subunits from U1, U2 snRNPs, decreases splicing fidelity switches gene output from linear to circRNAs in *Drosophila*.

1.4 Thesis aims

This work aims at understanding the role of RNAPII post-translational modifications in the production of circRNA during murine neuronal maturation. The functions of RNAPII modifications in co-transcriptional RNA processing have so far been addressed in several model organisms and mammalian cultured cells, but remain mostly unexplored in neurons. Among all eukaryotic cell types,

neurons display increased transcript complexity, such as high levels of alternative splicing and circRNA formation. Many alternative splicing events stem from neuron-specific RNA processing, but the contribution of RNAPII post-translational modifications is largely unexplored. I chose to focus my studies on circular RNA biogenesis because circRNAs are most abundant in neurons and evidence suggests that the crosstalk between transcription and splicing is important for their production. Nevertheless, the extent to which RNAPII modifications influence circRNA formation is still unknown.

To understand if the interplay between RNAPII and circRNA formation is similar or distinct between very different cells types and between neuronal subtypes, I explore this question in parallel in two neuronal differentiation timelines, from mESCs to mature dopaminergic neurons or spinal motor neurons. To study how specific patterns of RNAPII modifications relate to circRNA production, I explored transcriptome sequencing (total-RNA-seq) coupled with genome-wide occupancy by ChIP-seq of the spliceosome, RNAPII modifications, and transcription modulators (**Fig. 1.14**). I compared genes producing circRNAs (circ+) to genes not producing circRNAs in the differentiation systems used (circ-) to dissect the characteristics of circ+ genes at each time-point in both differentiation systems.

The first chapter of this thesis summarises the fields of transcription, splicing, and the interplay between both from the perspective of RNAPII CTD modifications. It also covers the current knowledge on circRNA function and biogenesis. The second chapter describes the experimental and computational approaches used in this work. The third chapter investigates the general features of circRNAs and genes producing circRNAs during neuronal differentiation. The fourth chapter addresses the contribution of RNAPII modifications and transcription regulation to circRNA formation in mESCs and throughout neuronal maturation. It also explores the effects of manipulating RNAPII release from the promoter on circRNA production. The fifth chapter discusses the results shown in this work and their implications in the context of transcription and circRNA fields.

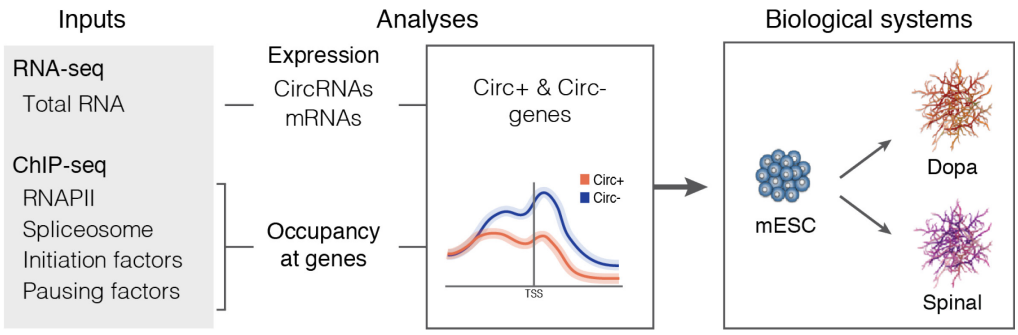

Figure 1.14 Approach to explore the role of RNAPII modifications in circRNA formation.
Taking advantage of several high-throughput sequencing datasets, I investigate the features of circRNA producing genes (circ+) compared with genes never producing circRNAs (circ-) in several stages of neuronal maturation.

Part II

Materials and methods

2 Materials and methods

2.1 Experimental procedures

2.1.1 mESC differentiation to dopaminergic neurons

All cell culture and differentiation procedures were performed by me, unless otherwise stated. RNA for biological replicate 1 from the dopaminergic neuron differentiation was provided by Dr. Carmelo Ferrai. RNA for biological replicate 2 from days 16 and 30 of the dopaminergic neuron differentiation was provided by Dr. Alexander Kukalev. Chromatin samples for ChIP-seq of NELF-E and U1C were provided by PhD. student Izabela Harabula (Pombo laboratory). ChIP-seq for RNAPII and mRNA-seq datasets from the *in vitro* differentiated dopaminergic neurons were obtained from published resources (Ferrai et al. 2017).

Mouse embryonic stem cells

Mouse ESCs from the 46C cell line (derived from E14tg2a, which express GFP under *Sox1* promoter (Ying et al. 2003), a kind gift from Prof. Domingos Henrique, were grown on gelatin-coated (0.1% v/v; Sigma cat# G1393) Nunc T25 flasks (ThermoScientific, cat# 21710025) in GMEM medium (Invitrogen, cat# 21710-082) supplemented with 10% Foetal Bovine Serum (Gibco, cat# 16141-079), 0.1 mM of β-mercaptoethanol (Gibco, cat# 31350-010), 2 mM L-glutamine (Gibco, cat# 25030-024), 1 mM sodium pyruvate (Gibco, cat# 11360070), 1% penicillin-streptomycin (Gibco, cat# 15140-122), 1% MEM Non-Essential Aminoacids (Gibco, cat# 11140-035) and 2000 U/ml of Leukemia Inhibitory factor (LIF; Millipore, cat# ESG1107). The medium was changed every day and cells were split every other day.

Prior to collection of material for ChIP and RNA, cells were grown for 48h in serum-free ESGRO Complete Clonal Grade Medium (Millipore, cat# SF001-500P)

supplemented with 1000 U/ml LIF, on gelatine coated Nunc 15 cm dishes. Medium was changed at 24h.

Early differentiation (mESC to day 3)

The protocol was developed by Abranches and colleagues (Abranches et al. 2009) and further optimized to larger scale cultures by C. Ferrai (Ferrai et al. 2017). Briefly, 46C mESCs were plated in high density (1.5x106 cells/ml) in serum-free ESGRO Complete Clonal Grade Medium (Millipore, cat# SF001-500P) supplemented with 1000 U/ml LIF, on gelatin-coated (0.1% v/v) Nunc 10 cm dishes (Gibco, cat# 150350). After 24h, cells were washed with PBS (Gibco, cat# 14040-133) and dissociated with 0.05% (v/v) trypsin (Gibco, cat# 25300-096) for 2 min at 37˚C. Cells were plated on 0.1% v/v gelatin-coated Nunc 10 cm dishes at a density of $1.6x10^6$ cells/dish in RHB-A medium (Takara-Clontech, cat# Y40001). Medium was changed every day, until day 3.

Dopaminergic neuron differentiation (days 16 and 30)

The protocol for dopaminergic neuron differentiation was based on the method developed by Jaeger and colleagues (Jaeger et al. 2011) and optimized by C. Ferrai (Ferrai et al. 2017; Fraser et al. 2015). Briefly, mouse EpiStem cell (EpiSC) clones were produced from 46C mESC after 4 weeks of growth in N2B27 basal medium supplemented with 20 ng/ml of Activin (R&D, Cat# 338-AC-050) and 12 ng/ml of FGF2 (Peprotech, cat# 100-18B). N2B27 basal medium is composed of 50% DMEM/F12 (Invitrogen, cat# 21331-020) and 50% Neurobasal medium (Invitrogen, cat# 21103-049), plus 0.5x N2 (Invitrogen, cat# 17502-048), 0.5x B27 (Invitrogen, cat# 12587-010), 0.05 M β-mercaptoethanol (Invitrogen, cat# 31350-010) and 2 mM L- glutamine (Invitrogen, cat# 25030-024). EpiSCs were grown in 6 well Nunc dishes (cat# 140675) coated with FBS (Gibo, cat# 10270) with N2B27 basal medium supplemented with Activin (R&D, cat# 338-AC-050) and FGF2 (Peprotech, cat# 100-18B).

The day before starting the differentiation, EpiSCs were plated on dishes coated with 15 ug/ml human plasma fibronectin (Millipore, cat# FC010), in N2B27 basal

medium supplemented with Activin and FGF2, to reach a confluence of 70-80% after 24h. The next day, differentiation was started by adding N2B27 basal medium supplemented with 1μM PD 0325901 (Axon, cat# 1408). On day 2, the cells were washed with PBS (Gibco, cat# 14040-133) and re-plated on Nunc 10-cm dishes (cat# 150350) coated with 15 μg/ml human plasma fibronectin (Millipore, cat# FC010) and cultured in N2B27 basal medium for 3 days. The N2B27 basal medium was changed daily. Afterwards, the medium was replaced with N2B27 basal medium supplemented with 100 ng/ml human Fibroblast Growth Factor 8 (hFGF8, Peprotech, cat# 100-25-25) and 200 ng/ml Sonic hedgehog (SHH, R&D, cat# 464-sh-025). After 4 days, the medium was replaced with N2B27 basal medium supplemented with 10 ng/ml Brain-Derived Neurotrophic Factor (BDNF, R&D, cat# 450-02-10), 10 ng/ml Glial cell-Derived Neurotrophic Factor (GDNF, R&D, cat# 450-10-10) and 200 μM L-ascorbic acid (Sigma-Aldrich, cat# A4544). Half of the medium volume was replaced every day with fresh medium until the day of sample collection.

2.1.2 mESC differentiation to spinal motor neurons

Spinal motor neuron differentiation until day 2 was developed by Prof. Esteban Mazzoni and colleagues (Mazzoni et al. 2013) and further extended to 25 days by PhD. student Disi An in the Mazzoni group(An et al. 2019). I have further adapted the protocol to larger cultures to be compatible with ChIP for RNAPII modifications. I learned the first part of the protocol from Dr. Silvia Velasco and received extensive advice for spinal motor neuron re-plating and extended culture from Disi An (both from the Mazzoni laboratory).

Embryoid body formation and induction of Spinal Motor neurons

mESCs bearing an inducible cassette for Ngn2, Isl1, and Lhx3 (NIL) factors were grown in 80/20 medium, i.e., 80% of 2i medium and 20% of mESC medium. 2i medium is composed of half Advanced DMEM/F12 (Invitrogen, cat# 12634010) and half Neurobasal medium (Invitrogen, cat# 10888-022), 1x N2 (Thermo-Fisher, cat# 17502-048), 1x B27 (Life technologies, cat# 17504044), 2 mM L-

glutamine (Invitrogen, cat# 25030-024), 0.1 mM of β-mercaptoethanol (Invitrogen, cat# 31350-010), 1000 U/ml of LIF, 3 μM of CHIR99021 (BioVision, cat# 1991-1) and 1μM of PD0325901 (Sigma, cat# pz0162). mESC medium is composed of Knockout™ DMEM (Gibco, cat# 10829-018) with 14% Foetal Bovine Serum (Gibco, cat# 16141-079), 0.1 mM of β-mercaptoethanol, 2 mM L-glutamine, 1x MEM Non-Essential Aminoacids (Gibco, cat# 11140-035), 1x Nucleosides (Millipore, cat# es-008-d) and 1000 U/ml LIF. NIL mESCs were grown on gelatin-coated (0.1% v/v) Nunc T25 flasks (136196) for about a week before differentiation. Medium was changed every day and cells were split every other day.

NIL mESCs differentiation is described in detail in (Mazzoni et al. 2013). Prior to differentiation, NIL cells were grown to form embryoid bodies (EBs). For this, NIL (Ngn2, Isl1 and Lhx3) mESCs at ~ 70-80% confluence were washed with PBS (Gibco, cat# 14040-133), dissociated with TryplE (Life technologies, cat# 12605-010) and incubated for 1 min at 37˚C. Cells were seeded at $3x10^6$ cells per 245mmx245mm Corning plate (cat# 431111) and grown in suspension in Differentiation medium. Differentiation medium is composed of half Advanced DMEM/F12 (Invitrogen, cat# 12634010) and half Neurobasal A medium (Invitrogen, cat# 10888-022), 10% of Knockout-SR serum (Invitrogen, cat# 10828028), 0.1 mM of β-mercaptoethanol and 2 mM L-glutamine (Invitrogen, cat# 25030-024). After 48h, EBs were collected and re-plated 1:2 in AK medium supplemented with 3 μg/ml of Doxycycline (Sigma, cat# d9891) to induce expression of NIL factors. After 48h, induced EBs were collected for total RNA-seq or ChIP-seq (day 2) or further dissociated and re-plated until day 8 of the spinal motor neuron differentiation.

Extended culture of Spinal Motor neurons

For extended culture of spinal motor neurons, induced EBs were collected and washed with PBS and dissociated with TryplE (Life technologies, cat# 12605-010) and incubated at room temperature for 1-1.5 min. After blocking of TryplE

with NIL mESC growth medium (which contains serum that blocks TryplE action), cells were washed in Differentiation medium and counted. Finally, cells were resuspended in Motor Neuron medium, composed of Neurobasal A medium (Invitrogen, cat# 10888-022), 2% Foetal Bovine Serum (Gibco, cat# 16141-079), 1x B27 (Life technologies, cat# 17504044), 0.5 mM L-glutamine (Invitrogen, cat# 25030-024) and 0.01 mM of β-mercaptoethanol supplemented with 3 µg/mL of Doxycycline (Sigma, cat# d9891), 4 µM of deoxyfluoruridine (FDU; Sigma, cat# F0503), 4 µM uridine (Sigma, U-3003), 10 ng/ml of Glial Derived Neurotrofic Factor (GDNF; R&D systems, cat# 512-GF-050), 10 ng/ml of Brain-Derived Neurotrophic Factor (BDNF; Peprotech, cat# 450-02), 10 ng/ml of Ciliary Neurotrophic Factor (CNTF; Peprotech, cat# 450-13-20), 10 µM Forskolin (Sigma, cat# F6886) and 100 mM 3-isobutyl-1-methylxanthine (IBMX; Sigma, cat# I5879). Motor neurons were plated in 15-cm Nunc dishes (cat# 168381) coated with 0.001% Poly-D-Lysine (Sigma, cat# P0899) overnight at 37˚C. Motor neurons were plated at $7x10^6$ cells/dish density and grown for 6 days with half of the medium refreshed every other day. After 6 days, cells were collected for ChIP or total RNA-seq (day 8).

2.1.3 Gene expression analyses

Single gene RNA expression

Total RNA was extracted with Trizol (Invitrogen, cat# 15596-018) according to manufacturer's instructions. Total RNA was treated with TURBO DNAse I (Ambion, cat# AM1907) in a 25 µl reaction according to manufacturer's instructions. 5 µg of total RNA were reverse transcribed with 50 ng random primers and 10 U of Superscript II reverse transcriptase (Invitrogen, cat# 18064-071) in a 20 µl reaction, as described in the protocol. cDNA was diluted 1:100 before quantifying the genes of interest by quantitative real-time PCR (qPCR). qPCR experiments were performed to confirm spinal motor neuron differentiation and NELF knockdown efficiency. qPCR was performed using SYBR No-Rox sensimix (Thermo Fisher, cat# S7563) in a 25 µl reaction and all primers were normalised to *Actb*. Primers spanning exon-exon junctions were used for

mRNA quantification. Primer quality was evaluated with a 2% agarose gel. Primer sequences are listed in **Table 2.1**. CircRNA expression was also confirmed for day 16 dopaminergic neurons, but qPCRs are not shown in this thesis.

Table 2.1 List of gene expression primers (FW – forward; RV - reverse)

Primer	Sequence (5' to 3')
Actb FW	TCTTTGCAGCTCCTTCGTTG
Actb RV	ACGATGGAGGGGAATACAGC
Pou5f1 FW	ACCTCAGGTTGGACTGGGCCT
Pou5f1 RV	GCCTCGAAGCGACAGATGGT
Hb9 FW	GAACACCAGTTCAAGCTCAACA
Hb9 RV	CTCTTCCGTCTTCTCCTCACTG
Chat FW	CGGTTTATTCTCTCCACCAG
Chat RV	TAACAGGCTCCATACCCATT
Nelf-A FW	GCACCGTGGACGAGATGA
Nelf -A RV	CAGAATGTCAGCAACCATGAGG
Spt5 FW	ATGCAGAGAAGATCAGGTCCTT
Spt5 RV	TGGAAATTGTGTTCCCGCTC

Total RNA-seq libraries

Total RNA was extracted and treated with TURBO DNAse I as above-mentioned. RNA quality was assessed by running a Agilent Bioanalyser 6000 RNA Nano assay (Agilent, cat# 5067-1511) and high quality samples were used for library preparation (RIN above 7.30 for all samples). 1 µg of DNAse-treated total RNA was used to produce total RNA-seq libraries with TruSeq Stranded Total RNA Sample preparation kit (Illumina, cat# RS-122- 2201). Total RNA-seq library quality was assessed before sequencing with Bioanalyser High Sensitivity DNA assay (Agilent, cat# 5067-4626). Libraries were sequenced with paired end sequencing in HiSeq2000 or NextSeq, following manufacturer's instructions.

2.1.4 Small interfering RNA (siRNA) treatments in mESCs

The siRNA oligos for NELF-A subunit and for SPT5 from Rahl and colleagues was used for knockdown of NELF (GCCATTTCTCTGGAGAGTTTA) and SPT5 (GCAATTTCCTTGATGCGGAAA) (Rahl et al. 2010). A second siRNA was designed with siRNA wizard tool (Invitrogen) to target SPT5 (GCGGAAATTCATCGCTTACCA). A control siRNA oligo with a scrambled sequence

was used as control (ATTCACCTAGGTTGATTCGTG) and designed with GenScript scrambled siRNA design tool (https://www.genscript.com/tools/create-scrambled-sequence). 46C mESCs were thawed and grown in standard conditions for a week before interference experiments. Transfection of was performed with 125 pmol of siRNA per $5x10^5$ cells per well of a 6-well Nunc plate (Thermo Fisher, cat# 140675) previously coated with gelatin (0.1% v/v; Sigma cat# G1393). All incubations were performed at room temperature. Briefly, each 125 pmol of siRNA was diluted in 250 µl of Opti-MEM reduced serum medium (Gibco, cat# 31985-070) and incubated for 20 min (siRNA mix). 5 µl/well of transfection reagent Lipofectamine-2000 (Invitrogen, cat# 11668-027) were incubated with 250 µl/well of Opti-MEM for 20 min (lipofectamine mix). Finally, the lipofectamine mix was mixed with the siRNA mix and incubated for 20 min. Meanwhile, 46C mESCs were split, washed, counted and resuspended in culture media. $5x10^5$ cells were plated per well and siRNA mix was added to the cells drop by drop. Cells were normally grown normally, with fresh media added every 24h. RNA samples were collected 72h after transfection and transfection efficiency was assessed by qPCR. RNA samples were subsequently used for total RNA-seq.

2.1.5 Chromatin immunoprecipitation (ChIP)

For plated cells, chromatin samples were produced as described in (Stock et al. 2007) and (Brookes et al. 2012), by crosslinking cells grown on plates in growth medium with 1% formaldehyde (Sigma, cat# F8775) for 10 min at 37°C, after which the reaction was quenched by adding glycine to a final concentration of 0.125 M. After fixation, cells were washed with ice-cold PBS (Gibco, cat# 14040-133) and placed in "swelling buffer" (25 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl (all from Sigma) and 0.1% NP-40 (Roche, cat# 11754599001) for 10 min, scraped from dishes and nuclei were isolated with a Dounce homogenizer (tight pestle) for 60 times followed by centrifugation. Nuclei were resuspended ($1x10^7$ nuclei/ml) in "sonication" buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na- deoxycholate and 0.1% SDS (all from Sigma))

and sonicated at 4˚C in a Diagenode Bioruptor. mESCs were sonicated at full power for 45 cycles of 30 s 'on' and 40 s 'off'. Dopaminergic and spinal motor neurons were sonicated for 60 cycles of 30 s 'on' and 30 s 'off'. Chromatin was centrifuged and, after disposal of the insoluble fraction, DNA content was quantified using alkaline lysis. Both swelling and sonication buffers were supplemented with phosphatase inhibitors 5 mM NaF (Sigma, cat# S7920) and 2 mM Na_3VO_4 (Sigma, cat# 450243) and with 1 mM PMSF (Sigma, cat# 78830) and protease inhibitor cocktail (Complete Mini EDTA-free, Roche, cat# 11836170001).

For induced embryoid bodies (day 2), chromatin preparation was performed as described in (Velasco et al. 2017). Briefly, embryoid bodies were collected and dissociated with TryplE (Life technologies, cat# 12605-010) for approximately 1 min. TryplE was blocked with NIL mESCs growth medium and cells were washed with 1x PBS. Cells were fixed with 1mM DSG (Alfa Aesar, cat# H58208-100mg) for 15 min at room temperature followed by fixing solution (50 mM HEPES-KOH pH 7.5, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 11% formaldehyde) for 15 min at room temperature. Fixation was quenched by adding glycine to a final concentration of 0.12 M. After fixation, cells were counted and cell pellets were frozen. Cells were thawed on ice, resuspended in 5 ml of Lysis Buffer A (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol (v/v), 0.5% Igepal (v/v), 0.25% Triton X-100 (v/v)) and incubated at 4˚C for 10 min in rotation. Samples were spun down for 5 min at 1.350 g, resuspended in 5 ml Lysis Buffer B (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0) and incubated for 10 min at 4°C in rotation. Samples were spun down for 5 min at 1.350 g and resuspended in 3 ml of Sonication Buffer (50 mM HEPES pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton X-100, 0.1% sodium deoxycholate (w/v), 0.1% SDS (w/v)). Chromatin was sonicated (Diagenode Bioruptor) at full power for 60 cycles of 30 s 'on' and 30 s 'off'.

After chromatin was extracted, immunoprecipitation was performed with several antibodies listed on **Table 2.2**.

Table 2.2 List of antibodies used for ChIP

Antibody	Raised in	Clone	Amount (µg)	Source	Catalogue number
RNAPII-S5p	Mouse (IgG)	CTD4H8	10	Covance	MMS-128P
RNAPII-S7p	Rat (IgG)	4E12	10	Kind gift from Dirk Eick	-
RNAPII-S2p	Mouse (IgM)	H5	10	Covance	MMS-129R
NELF E	Rabbit	ab170104	4	Abcam	ab170104
U1C	Rat (IgG)	4H12	10	Santa Cruz Biotechnology	sc-101549
Digoxigenin	Mouse (IgG)	-	10	Jackson Lab	200-002-156
Rabbit anti-mouse (IgG+IgM)	Rabbit	-	10	Jackson Lab	315-005-048
Rabbit anti-mouse (IgM)	Rabbit	-	10	Jackson Lab	315-005-020
Rabbit Anti-Rat IgG	Rabbit	-	10	Jackson Lab	312-005-045

For mouse anti-RNAPII S5p, rat anti-RNAPII S7p, mouse anti-RNAPII S2p, and mouse anti-Digoxigenin, 50 µl of protein G magnetic beads (Active Motif, cat# 53014) were incubated with 10 µl of rabbit anti-mouse (IgG+IgM) bridging antibodies for 1h at 4˚C and washed with sonication buffer. For rat anti-U1C, beads were incubated with rat anti-mouse (IgG) bridging antibodies. For rabbit anti-NELF E antibody no bridging antibody was necessary as beads affinity for rabbit is very high, and beads were incubated for 1h at 4˚C. For immunoprecipitation, 700 µg of chromatin with specific antibodies and corresponding controls were incubated with beads, overnight at 4˚C. After immunoprecipitation, beads were washed as previously described (Brookes et al. 2012; Stock et al. 2007). Immunoprecipitated complexes were eluted from beads with 50 mM Tris-HCl pH 8.0, 1 mM EDTA and 1% SDS, 5 min at 65˚C and 15 min at room temperature in rotation. Reverse-crosslinking was performed after addition of NaCl to a final concentration of 155 mM and 10 µg RNAse A (Sigma,

cat# R4642), overnight (for qPCR) or for 8h (for ChIP-seq libraries). Afterwards, samples were incubated with 100 µg of proteinase K (Roche, cat#3115836001) and EDTA with a final concentration of 5 mM, for 2h at 50˚C. DNA was extracted by phenol-chloroform (Millipore, cat# KP31757), precipitated with ethanol in the presence of 20 ng/ml of glycogen (Invitrogen, cat# AM9510) and eluted in 22 µl of TE buffer (for qPCR) or water (for libraries).

DNA concentrations were determined by Quant-iT PicoGreen dsDNA fluorimetric assay (Invitrogen, cat# P7589) and samples were diluted to a final concentration of 0.2 ng/µl. Immunoprecipitated and input samples (0.5 ng each) were analyzed by qPCR using SYBR No-Rox sensimix (Bioline). Quantitative PCR "cycle over threshold" (Ct) values from immunoprecipitated samples (RNAPII or control antibody) were subtracted from the input Ct values and the fold enrichment over input was calculated using the equation 2(input Ct – IP Ct). qPCRs for validation of ChIP experiments were performed but are not shown in this thesis.

For ChIP-seq library production the TruSeq ChIP Sample Preparation kit (Illumina cat# IP-202-1012) was used with minor changes. 5 to 10 ng of immunoprecipitated DNA were used and protocol was performed as described by the manufacturer until adapter ligation. After this step, DNA fragment enrichment was performed as described by the manufacturer and DNA was purified on a 2% ultra-pure agarose gel (Invitrogen, cat# 16500500). DNA fragments between 250 and 600 bp were selected and DNA was purified with the MinElute Gel Extraction Kit (Qiagen, cat# 28604). Before sequencing, quality control of the libraries was assessed with Bioanalyser High Sensitivity DNA assay (Agilent, cat# 5067-4626). Libraries were sequenced with single end sequencing with NextSeq (Illumina), following manufacturer's instructions.

2.1.6 Immunofluorescence

Spinal motor neurons – embryoid bodies

Cells were processed for immunofluorescence as described in (Mazzoni et al. 2013). Briefly, EBs were fixed with 4% (v/v) paraformaldehyde (PFA; VWR, cat#

43368.9M) in PBS (Gibco, cat# 14040-133), embedded in OCT compound embedding medium (Fisher, cat# 23-730-571) and sectioned (15-20 μm thick) for staining. Cells were permeabilized with 0.2% Triton T-100 (in PBS) and blocked with 10% FBS (Gibo, cat# 10270) in PBS for 30 min at room temperature. Primary antibodies were incubated overnight at 4˚C. Cells were then washed 3 times for 10 min each with PBS, before incubation with secondary antibodies for 1h at room temperature. Cells were washed twice for 10 min in PBS, at room temperature, and stained with 0.5 μg/mL 4′,6-diamidino-2-phenylindole (DAPI; Sigma, cat# D9542-5MG) in PBS for 2 min. Cells were washed once with PBS for 10 min and slides were dried for 10 min at 37˚C. Finally, cells were mounted with Mowiol 4-88 (Sigma, cat# 81381) and dried overnight. Images were acquired on a confocal laser scanning microscope (Leica TCS SP8; 63x oil objective, NA 1.4), with a pinhole equivalent to 1 Airy disk. Images from different channels were collected sequentially to prevent fluorescence bleed-through. Raw images (TIFF files) were merged in ImageJ and contrast stretched without thresholding in Adobe Photoshop. Details of antibodies used are shown in Table 2.3.

Spinal motor neurons – plated neurons

Volumes are for 12-well plates. All incubations are at room temperature unless otherwise specified. Spinal motor neurons were briefly washed with 500 μl of 4% (v/v) PFA (VWR, cat# 43368.9M) in 125 mM HEPES-NaOH, pH 7.4. Cells were re-fixed, first with 4% PFA in 125 mM HEPES, for 10 min, and after with 8% PFA in 125 mM HEPES, for 10 min, all at room temperature. Cells were double fixed to preserve neuronal processes better. Cells were then stored in 1 ml PBS supplemented with 0.05% NaN_3 (Sigma, cat# S2002). For staining, cells were permeabilized for 10 min with 0.2% (v/v) Triton X-100 (Sigma, cat# T9284) in PBS, and blocked with 10% goat serum (Jackson lab, 005-000-121) in PBS for 30 min. Primary antibodies incubated overnight at 4˚C. Cells were washed 3 times for 10 min each, in 1 ml PBS. After washes, secondary antibody was added and incubated for 1h at room temperature, followed by two washes in PBS for 10 min.

Cells were then stained with 0.5 μg/mL DAPI (Sigma, cat# D9542-5MG) for 2 min. Coverslips were dried for 10 min at 37˚C. Finally, coverslips were mounted in Mowiol 4-88 (Sigma, cat# 81381) and imaged as described above. Details of antibodies used are shown in **Table 2.3**.

Table 2.3 List of antibodies used for immunofluorescence

Primary antibody	Raised in	Dilution 1ry	Source	Secondary antibody	Source	Dilution 2ry
β-tubulin 3	Rabbit	1:10000	Sigma-Aldrich (T2200)	Alexa Fluor 546 Goat Anti-Rabbit IgG (H+L)	Invitrogen (A-11010)	1:1000
V5	Mouse	1:1000	Invitrogen (R96025)	Alexa Fluor 488 Goat Anti-Mouse IgG (H+L)	Invitrogen (A-11001)	1:1000
Hb9	Mouse	1:500	Developmental Studies Hybridoma Bank (81.5C10)	Alexa Fluor 488 Goat Anti-Mouse IgG (H+L)	Invitrogen (A-11001)	1:1000
Isl1/2	Mouse	1:500	Developmental Studies Hybridoma Bank (39.4D5)	Alexa Fluor 488 Goat Anti-Mouse IgG (H+L)	Invitrogen (A-11001)	1:1000

2.2 Computational approaches

All computational analyses were performed by me, unless otherwise stated.

2.2.1 Bulk RNA-seq analyses

Published data were downloaded from the GEO repository via direct link or using the SRAtoolkit. RNA-seq processing steps were previously tested and implemented in the Pombo group.

Quality control was performed on sequencing data (.fastq files) before further processing using FastQC software (Andrews 2010). Sequenced reads were aligned to the mouse genome annotation (assembly mm9, Ensembl build 67). Total RNA-seq libraries were mapped with STAR 2.5 (Dobin et al. 2013) with the following parameters:

```
--runThreadN 6 --quantMode TranscriptomeSAM
```

Mapped bam files were used to quantify gene expression in RSEM 1.3.1 (Li and Dewey 2011) with default parameters and Ensembl release 67 (equivalent to mm9) as reference transcriptome. To define transcript isoforms, Ensembl ensGene table was used (mm9). Transcript isoforms per time-point were selected according to the several parameters in the following order: 1) highest level of expression; 2) canonical isoform (longest length); 3) first isoform of the gene. Gene expression was calculated with Transcript Reads per Million (TPM), and, in some instances, TPM values were represented as Log10 TPM. Prior to calculating Log 10, a pseudo-count was added (as stated in the graph).

For visualisation, UCSC Genome Browser (http://genome.ucsc.edu) was used. All viewing settings were set to default, apart from "Windowing function", which was set to "Mean", "Smoothing Window", set to "2 pixels", and inclusion of zero on y-axis.

Biological replicates were treated separately and list of RNA-seq datasets are shown in **Tables 2.4** and **2.5**.

Table 2.4 Details of published RNA-seq datasets used in this work

Library	Sample	Biological replicate	GEO	Publication
poly(A) selected RNA	46C mESC	1	GSE94364	(Ferrai et al. 2017)
poly(A) selected RNA	Dopa differentiation day 1	1	GSE94364	(Ferrai et al. 2017)
poly(A) selected RNA	Dopa differentiation day 3	1	GSE94364	(Ferrai et al. 2017)
poly(A) selected RNA	Dopa differentiation day 16	1	GSE94364	(Ferrai et al. 2017)
poly(A) selected RNA	Dopa differentiation day 30	1	GSE94364	(Ferrai et al. 2017)

Table 2.5 Details of total RNA-seq datasets produced in this work

Library	Sample	Biological replicate	# sequenced reads	# mapped reads	% mapped reads	Sequencing details
total RNA	46C mESC	1	327802420	310908560	94.8	paired end, 100 bp
total RNA	Dopa differentiation day 1	1	287425748	266880166	92.9	paired end, 100 bp
total RNA	Dopa differentiation day 3	1	309893252	295179964	95.3	paired end, 100 bp
total RNA	Dopa differentiation day 16	1	387039624	367504714	95.0	paired end, 100 bp
total RNA	Dopa differentiation day 30	1	405007200	388369278	95.9	paired end, 100 bp
total RNA	Spinal differentiation day 2	1	529870450	507342730	95.7	paired end, 75 bp
total RNA	Spinal differentiation day 8	1	529185526	512444890	96.8	paired end, 75 bp
total RNA	46C mESC	2	374944792	358534850	95.6	paired end, 75 bp
total RNA	Dopa differentiation day 1	2	347515932	330375434	95.1	paired end, 75 bp
total RNA	Dopa differentiation day 3	2	441595702	423624621	95.9	paired end, 75 bp
total RNA	Dopa differentiation day 16	2	407660384	397393204	97.5	paired end, 75 bp
total RNA	Dopa differentiation day 30	2	380493256	372423182	97.9	paired end, 75 bp
total RNA	Spinal differentiation day 2	2	471523412	457992609	97.1	paired end, 75 bp
total RNA	Spinal differentiation day 8	2	752320392	738070997	98.1	paired end, 75 bp
total RNA	46C mESC siRNA Scrambled	1	491450268	459846057	93.6	paired end, 75 bp
total RNA	46C mESC siRNA NELF-A	1	664990868	624716230	93.9	paired end, 75 bp
total RNA	46C mESC siRNA Scrambled	2	592789958	552156735	93.1	paired end, 75 bp
total RNA	46C mESC siRNA NELF-A	2	575459038	533026033	92.6	paired end, 75 bp

2.2.2 Bulk ChIP-seq analyses

Published ChIP-seq data were downloaded from the GEO repository via direct link or using the SRAtoolkit. ChIP-seq processing steps were previously tested and implemented in the Pombo group.

QC was performed on sequencing data (.fastq files) before further processing using FastQC software (Andrews 2010). Sequenced reads were aligned to the mouse genome annotation (assembly mm9) using Bowtie2 (Langmead and Salzberg 2012) with default parameters. For each bam file generated a bed file was generated with the Samtools software (Li and Durbin 2009). Duplicate reads occurring more often than a given threshold were removed. This threshold was calculated for each dataset individually as the 95th percentile of the frequency of read distribution, with a script developed by Dr. Inês de Santiago (Pombo lab). Lists of unpublished and published ChIP-seq datasets are shown in **Tables 2.6** and **2.7**, respectively.

Table 2.6 Details of ChIP-seq datasets produced in this work

Library	Sample	Antibody	# sequenced reads	# mapped reads	% mapped reads	Sequencing details
RNAPII S5p	46C mESC	CTD4H8	175549777	170324490	97.0	single end, 75bp
	Spinal day 2	CTD4H8	74723846	72824121	97.5	single end, 75bp
	Spinal day 8	CTD4H8	61151330	59328371	97.0	single end, 75bp
RNAPII S7p	46C mESC	4E12	88297762	85518036	96.9	single end, 75bp
RNAPII S2p	46C mESC	H5	148969230	143866173	96.6	single end, 75bp
U1C	46C mESC	sc-101549	62762449	60537495	96.5	single end, 75bp
NELF-E	46C mESC	ab170104	48898842	43232655	88.4	single end, 75bp
	Dopa day 16	ab170104	40597445	25772622	63.5	single end, 75bp
Digoxigenin	46C mESC	200-002-156	49173922	44571428	90.6	single end, 75bp
Mock	46C mESC	-	90054263	85008960	94.4	single end, 75bp

Table 2.7 Details of published ChIP-seq datasets used in this work

Dataset	Sample	Antibody	GSE	Publication
TAF1	D3 mESC	Produced by authors	GSE31270	(Liu et al. 2011)
TBP	D3 mESC	Anti-TBP (ab62126)	GSE31270	(Liu et al. 2011)
NELF-A	V6.5 mESC	A-20 (sc-23599)	GSE20530	(Rahl et al. 2010)
CDK7	V6.5 mESC	A300-405A	GSE60027	Young lab
CDK8	V6.5 mESC	sc-1521	GSE60027	Young lab
CDK9	V6.5 mESC	H-169 (sc-8338), C-20 (sc-484)	GSE44288	(Whyte et al. 2013)
Total RNAPII	V6.5 mESC	PolII N20 sc-899	GSE20530	Young lab
RNAPII S5p	Dopa day 16	CTD4H8	GSE94364	(Ferrai et al. 2017)
	Dopa day 30	CTD4H8	GSE94364	(Ferrai et al. 2017)
RNAPII S7p	Dopa day 16	4E12	GSE94364	(Ferrai et al. 2017)
	Dopa day 30	4E12	GSE94364	(Ferrai et al. 2017)
Digoxigenin	Dopa day 16	200-002-156	GSE94364	(Ferrai et al. 2017)

Calculation of ChIP-seq coverage and enrichment level

ChIP-seq coverage at a given window was determined using bedtools coverage with -d parameter, from BEDTools suite v2.17.0 (Quinlan and Hall, 2010), and a custom python script developed by Dr. Tiago Rito (previously Pombo group) which calculated coverage per bp. Further calculation of coverage per 10 bp and generation of average profiles were produced with a custom R scripts I developed.

Amount of ChIP-seq enrichment was calculated using bedtools coverage with -counts parameter, which reports the total number of reads at a given window. Counts were imported to R and average counts were calculated (total counts/window length). ChIP-seq signal mount at TSS was determined for a $\pm$500 bp window centered on the TSS, E1-I1 border spanned -50 bp of E1 end to 200 bp after, and TES amount was calculated from TES to 2 kb after.

Determining positive peaks

Positive peaks for all ChIP-seq datasets were identified Bayesian Change Point (BCP) peak finder (Xing et al. 2012) with Histone Modification mode. Dig dataset was used as control for RNAPII S5p, S7p, and U1C. Mock dataset was used as control for RNAPII S2p and NELF-E. Positive peaks were further intersected with windows at the TSS ($\pm$ 1000 bp) for all genes using bedtools intersect. Positive peaks at promoters were used to calculate overlaps and correlation of enrichment levels between ChIP-seq datasets. Density of ChIP-seq reads at TSSs was determined with R.

2.2.3 CircRNA identification and processing

CircRNA identification was performed by Petar Glažar (Nikolaus Rajewsky group, BIMSB, MDC, Berlin). CircRNAs were identified with find_circ pipeline (Memczak et al. 2013). Briefly, this pipeline filters out reads that do not align contiguously to the genome, retaining spliced reads. Then, the terminal parts of the reads are mapped to find unique anchor positions. The anchor alignments are extended such that the original read must align to the genome and its breakpoint is flanked

by GU/AG, the typical splice site junction sequence. Reads in inverted orientation would represent a circRNA event. An illustration of the find_circ pipeline is shown in **Fig. 2.1**. The output files of the pipeline consist in two files with circRNA annotation and quantification.

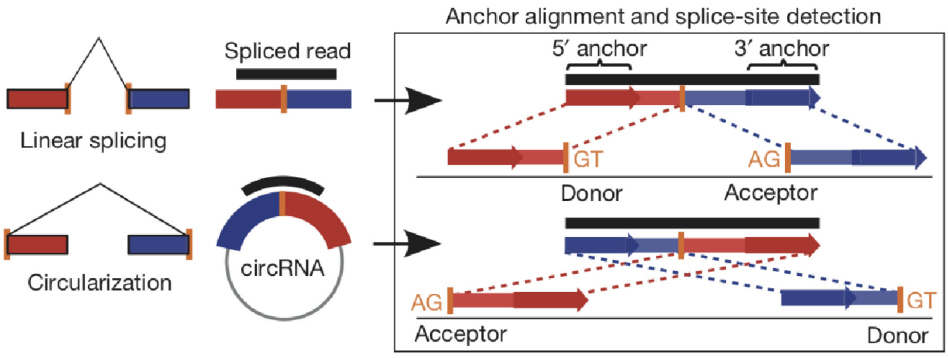

Figure 2.1 Schematic representation of find_circ pipeline.
Exons are represented in red and blue. Black lines represent reads spanning spliced or back-spliced junctions (Memczak et al. 2013).

After merging both files, I recomputed circ-to-linear ratios by adding a pseudo count to the linearly spliced reads to avoid dividing by zero. To calculate circRNA start and end positions in the transcript and number of exons in the circRNA, the start and end position of all circRNAs were intersected with the position of exons in the corresponding transcript using bedtools intersect and a personal script.

Gene coordinates and features

Whenever analyses are respective to circular RNAs, gene coordinates and features correspond to transcript isoforms attributed directly to circular RNAs by the find_circ pipeline. Whenever analyses are respective to genes producing circRNAs, gene coordinates and features correspond to the transcript isoform selected per time-point.

CircRNA expression per gene

To determine circRNA expression per gene ($CircRP_{100}M$), the total number of back-spliced reads per gene were counted and normalized to the number of circRNAs produced per gene. This value was further normalized to the total

number of mapped reads to compare circRNA expression between time-points. A scaling factor of 10^8 was used for more readable values. The formula applied was the following:

$$\text{Back-spliced reads per gene (CircRP}_{100}\text{M)} = \frac{\text{Back-spliced reads / Number circRNAs}}{\text{Total mapped reads / } 10^8}$$

2.2.4 Plot generation

Heatmaps were generated using the R package *pheatmap* (Kolde 2015). Bar plots, dot plots boxplots and violin plots were produced with Excel or R, using R generic functions or the package *ggplot2* (Wickham 2016).

2.2.5 Gene Ontology enrichment analyses

Gene Ontology (GO) enrichment analyses were performed using GO-Elite (version 1.2.5, Gladstone Institutes; http://genmapp.org/go_elite). Default parameters were used as filters: z-score threshold > 1.96, permutation-derived p-value < 0.05, number of genes changed > 2. Pruned results are reported, to reduce term redundancy. Over-representation analysis was performed with the "permute p-value" option, 2000 permutations. The group of genes used as background is specified in the corresponding figure legend.

2.2.6 Statistical analyses

Statistical analyses were performed with R software and the statistical test used is stated in each figure legend.

Part III

Results

3 CircRNA expression in dopaminergic and spinal motor neuron differentiation

3.1 Research motivation and aims

CircRNA production is cell-type specific, very dynamic, and occurs at high levels in neuronal cells (Rybak-Wolf et al. 2015). Evidence also points to circRNA production being a highly regulated process that may be intimately linked to co-transcriptional splicing regulation: genes producing circRNAs are transcribed by a fast elongating RNAPII and, in some instances, circRNA production is associated with read-through transcription events and decreased levels of spliceosome components (Ashwal-Fluss et al. 2014; Liang et al. 2017; Zhang et al. 2016). Numerous studies have shown that RNAPII post-translational modifications, namely S2p, S5p, and S7p, play a fundamental role in the co-transcriptional recruitment of different protein complexes to process nascent RNA during the transcription cycle (Brookes and Pombo 2009; Zaborowska, Egloff, and Murphy 2016). However, the contribution of specific RNAPII post-translational modifications to circRNA formation is still unknown. Additionally, circRNA production may be regulated differently in distinct cell types; for example in mESCs and differentiated neurons, or even between neuronal subtypes.

To address these questions, I started by quantifying circRNA expression and exploring the features of circRNAs and genes producing circRNAs in two neuronal differentiation systems from mESCs. The first differentiation system captures mESC exit from pluripotency, as well as immature and mature dopaminergic neurons (Abranches et al. 2009; Ferrai et al. 2017; Fraser et al. 2015; Jaeger et al. 2011). The second differentiation system directly programs mESCs to spinal motor neurons and was developed by Esteban Mazzoni and colleagues (Mazzoni et al. 2013) and further extended by PhD. student Disi An in the Mazzoni group (An et al. 2019). I chose to focus on dopaminergic and spinal motor neurons

because these have very different biological functions. Dopaminergic neurons are mostly found in the ventral midbrain, use the neurotransmitter dopamine, are involved in emotion-based behaviours, such as reward and motivation, and are implicated in addiction. Loss of dopaminergic neurons also leads to neurological disorders, for example Parkinson's disease or schizophrenia (Chinta and Andersen 2005). Conversely, spinal motor neurons are cholinergic cells which transmit motor information through very long axons from the brain to effector targets, such as muscles. Spinal motor neurons are also relevant in disease; their progressive degeneration is the underlying cause of amyotrophic lateral sclerosis and spinal muscular atrophy (Davis-Dusenbery et al. 2014).

In this chapter, I characterize the features of circRNAs and genes producing circRNAs during dopaminergic and spinal motor neuron maturation. I produced total RNA-seq libraries from mESCs and 4 time-points along the dopaminergic neuron differentiation. I also established spinal motor neuron differentiation in the Pombo group and produced total RNA-seq libraries. These datasets were subsequently used for circRNA and gene expression quantification. Finally, I characterized circRNA expression dynamics during neuronal maturation and explored the features of genes producing circRNAs.

3.2 Contribution disclosure

From Pombo group: Carmelo Ferrai provided RNA samples and produced total RNA-seq library production for biological replicate 1 from the dopaminergic neuron differentiation (mESCs to day 30). Giulia Caglio mapped the total RNA-seq datasets for biological replicate 1 from the dopaminergic neuron differentiation (mESCs to day 30) after sequencing. Alexander Kukalev provided RNA samples for biological replicate 2 from the dopaminergic neuron differentiation (days 16 and 30). Markus Schueler contributed to the calculation of the number of circRNAs per gene. Tiago Rito devised the strategy for normalization of circularized reads per gene.

From Nikolaus Rajewsky group, MDC: Petar Glažar performed circRNA identification together with a list with matched linear transcripts and performed corresponding quality controls for the dopaminergic and spinal motor neuron differentiation.

From Esteban Mazzoni group, NYU: Silvia Velasco taught me the spinal motor neuron differentiation and Disi An advised on spinal motor neuron re-plating.

3.3 Notes to the reader

Results shown in this chapter correspond to biological replicate number 1, unless otherwise specified. All analyses were performed for biological replicate 2 in parallel and were consistent with biological replicate 1. Analyses shown in this chapter considered only circRNAs robustly detected in two biological replicates, unless otherwise specified.

3.4 Dopaminergic neuron differentiation

To explore features of circRNA expression during dopaminergic neuron differentiation, I took advantage of a differentiation system established in the Pombo lab by Dr. Carmelo Ferrai (Ferrai et al. 2017; Fraser et al. 2015). The two differentiation systems capture different stages of mESCs differentiation to the neuronal lineage, yielding highly homogenous cell populations (**Fig. 3.1**). The first protocol was developed by Abranches and colleagues and recapitulates the early exit of mESC from pluripotency towards a neuronal fate (days 1 and 3) (Abranches et al. 2009). mESCs grown in feeder-free conditions are differentiated in monolayer towards an early neuronal precursor (NPC) fate, through endogenous synthesis of FGF and Notch which is stimulated by growth in the absence of serum, in a synthetic medium that lacks BMP signals (inhibitory to the neuronal fate). The second protocol was developed by (Jaeger et al. 2011) and differentiates mESCs to mature dopaminergic neurons, which are functionally active (Ferrai et al. 2017). In this second approach, mESCs are first differentiated into stable EpiSC clones, which are primed for differentiation. EpiSCs are then

differentiated in monolayer by blocking FGF/ERK signalling, which induces a neuroectodermal fate. This is followed by the activation of FGF signalling in combination with SHH stimulation to promote fate specification and differentiation. At day 16, dopaminergic neurons are post-mitotic and express the pan-neuronal marker β-tubulin III (TUBB3). These neurons also express the dopaminergic neuron marker Tyrosine Hydroxylase (TH) at low levels, confirming the dopaminergic fate, but electrophysiological recordings show that these cells display an immature behaviour (Ferrai et al. 2017). Dopaminergic neurons reach maturity at day 30, with increased TH expression and electrophysiological activity characteristic of mature dopaminergic neurons (Ferrai et al. 2017).

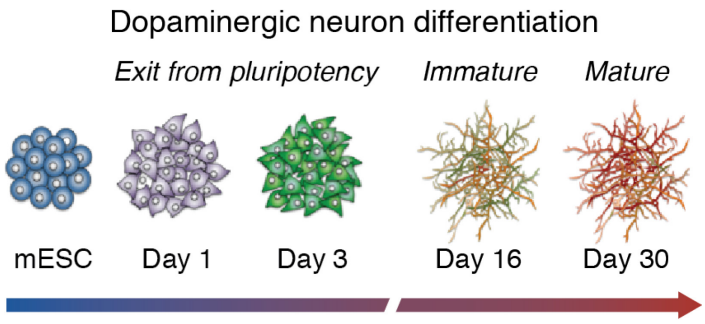

Figure 3.1 Overview of dopaminergic neuron differentiation.
Schematic summary of mESC differentiation to mature dopaminergic neurons. Days 1 and 3 recapitulate exit from pluripotency towards a neuronal progenitor fate and days 16 and 30 correspond to immature and mature dopaminergic neurons. Adapted from (Ferrai et al. 2017).

To study circRNA expression dynamics throughout dopaminergic differentiation, we produced total RNA-seq datasets from RNA extracted from two biological replicates of the differentiation from mESCs to the 4 stages into dopaminergic neurons (mESCs, days 1, 3, 16 and 30). Gene expression measured from total RNA-seq datasets (TPM) from the two biological replicates correlate well for all time-points analyzed (Pearson's 0.91-0.93, **Fig. 3.2A**) and expression of marker genes for different differentiation stages was as expected (**Fig. 3.2B**) (Abranches et al. 2009; Ferrai et al. 2017). *Nanog*, a homeobox transcription factor essential for stem cell renewal, peaks at the mESC stage and is lowly expressed or undetected after day 1. *Pou5f1*, a POU-containing homeobox transcription factor

important for pluripotency, is also highly expressed in mESCs and strongly decreased by day 3. Early differentiation marker *Fgf5* is transiently expressed from days 1 to 3 and neuronal markers *Hes5* and *Blbp* are expressed from day 3 onwards (Abranches et al. 2009; Ferrai et al. 2017). The pan-neuronal marker *Tubb3* is expressed from day 16 onwards, whereas *Th* expression is highest at day 30 (Jaeger et al. 2011; Ferrai et al. 2017). After careful validation of marker genes expression, total RNA-seq datasets were used for circRNA identification and quantification.

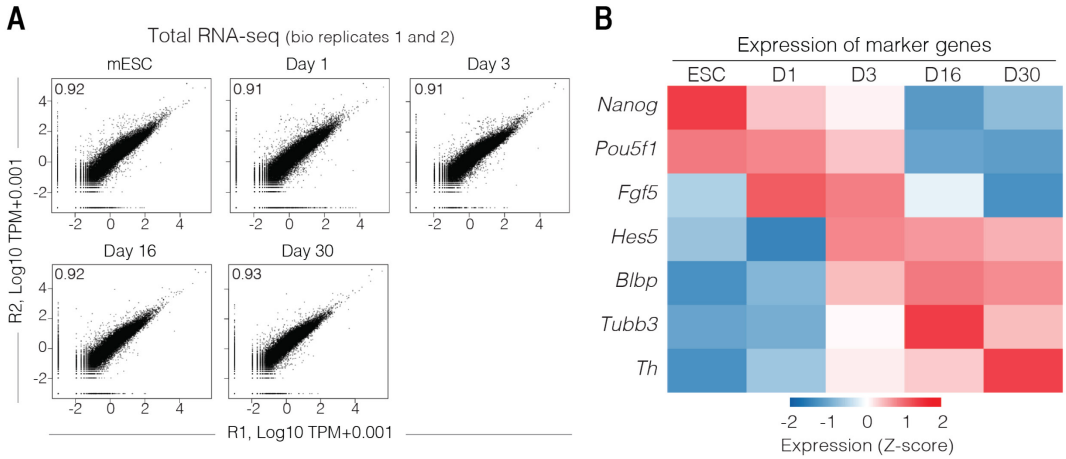

Figure 3.2 Marker gene expression during dopaminergic neuron maturation.
A) Pearson's correlation of gene expression for 2 biological replicates in mESCs and 4 time-points during neuronal differentiation. **B)** Expression of marker genes during neuronal differentiation into dopaminergic neurons. Gene expression is depicted as Z-score.

3.5 Spinal motor neuron differentiation

To explore features of circRNA expression during spinal motor neuron differentiation, I started by establishing the differentiation system developed by Esteban Mazzoni (Mazzoni et al. 2013; Velasco et al. 2017), which directly programs mESCs towards the spinal motor neuron fate in a very fast and efficient manner. This approach uses a mESC cell line that harbors a polycistronic doxycycline-inducible expression cassette with three transcription factors, *Ngn2*, *Isl1* and *Lhx3* (NIL) separated by 2A peptides (**Fig. 3.3A**, left panel). Firstly, EBs are formed and doxycycline is added after 2 days to the growth medium. Two days after induction, EBs are dissociated, re-plated and grown for 6 days in a supplement cocktail that mimics trophic factors released by supporting cells,

such as glia, to achieve increased neuronal maturity **(Fig. 3.3A**, right panel) (An et al. 2019). Two days after induction, cells display spinal motor neuron features, such as expression of the spinal motor neuron marker Motor Neuron And Pancreas Homeobox 1 (HB9) and the pan-neuronal marker TUBB3. These cells have properties of immature neurons, but become electrophysiologically functional when cultured on astrocytes for 7 days and exhibit robust outgrowth of axons exiting the spinal cord when implanted in the brain of chick embryos similarly to *in vivo* spinal motor neurons (Mazzoni et al. 2013). After induction, EBs are dissociated, re-plated and grown for 6 days to achieve increased maturity (An et al. 2019).

qPCR and immunofluorescence imaging representative of three biological replicates of the spinal motor neuron differentiation show that I successfully established the differentiation protocol. RNA expression of marker genes by qPCR shows pluripotency marker *Pou5f1* is reduced in induced EBs (D2) when compared to control EBs (CTL) *i.e.* EBs grown for 2 days without doxycycline which should not express NIL factors (**Fig 3.3B**). Conversely, Hb9 expression is higher in induced EBs (D2) compared with control EBs (CTL), which is also decreased in plated neurons (D8). The cholinergic neuron marker *Chat* is detected at day 2 and is expressed at the highest level at day 8, confirming spinal motor neuron identity. Results from immunofluorescence imaging corroborate RNA expression. Two days after adding doxycycline (day 2), induced EBs show increased expression of the spinal motor neuron marker HB9 and TUBB3 when compared to control EBs (**Fig. 3.3C**, control vs induced EBs). An antibody against the V5 tag confirms that the expression cassette was induced in EBs treated with doxycycline (**Fig. 3.3C**, control vs induced EBs). All neuronal and induction markers are absent in control EBs. After induction, EBs were dissociated and re-plated for six days (day 8), in the absence of supporting cells. Plated neurons display decreased HB9 expression and sustained TUBB3 expression (**Fig. 3.3C**, plated neurons) (Mazzoni et al. 2013; An et al. 2019; Davis-Dusenbery et al. 2014).

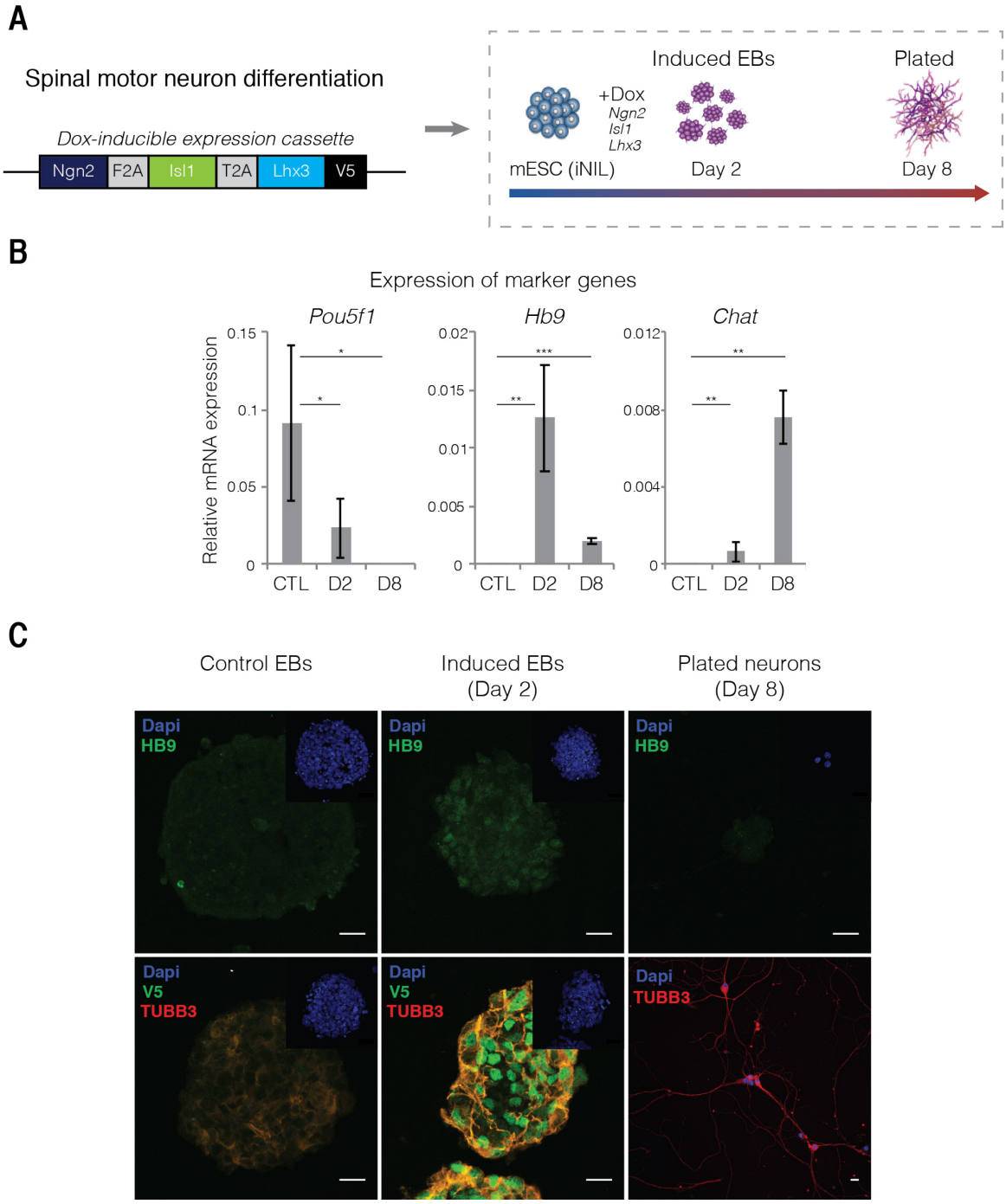

Figure 3.3 Overview of spinal motor neuron differentiation.
A) Schematic summary of direct programming of iNIL mESCs to spinal motor neurons. Top panel represents the doxycycline-inducible cassette expressing NIL factors. Bottom panel depicts the key stages in the spinal motor neuron differentiation protocol (Mazzoni et al. 2013). **B)** RNA expression of marker genes by qPCR in control EBs which were not induced by dox (CTL), induced EBs (D2), and plated neurons (D8). Relative levels are normalised to the *Actb* RNA. Mean and standard deviation from three biological replicates. Unpaired t-test, * p-value < 0.05, *** p-value < 0.01, *** p-value < 0.001. **C)** Immunofluorescence of expression of markers in control and induced EBs (day 2) and plated spinal motor neurons (day 8). β-tubulin III is pseudo-colored in red and HB9 or V5 in green. Nuclei are stained with DAPI. Scale bar, 20 μm.

To evaluate gene expression during spinal motor neuron differentiation, I collected RNA samples from induced EBs and plated spinal motor neurons (days 2 and 8, respectively) and produced total RNA-seq datasets from two biological replicates. Gene expression for both biological replicates correlates very well (Pearson's 0.91-0.96) and marker genes are expressed accordingly (**Fig. 3.4A**).

Next, I checked for the expression of marker genes using as comparison the total RNA-seq from mESC-46C (**Fig. 3.4B**). It was not useful to analyze the mESC (iNIL) genome-wide, since the studies presented here aimed at exploring distinct cell states. The pluripotency marker *Pou5f1* again shows decreased expression in day 2 and is undetected by day 8 compared with its expression in mESC-46C, whereas expression levels of NIL factors are abundant in induced EBs at days 2 and 8. *Tubb3* is also detected from day 2 onwards. *Hb9* gene expression peaks at day 2 and *Chat* is detected at day 2 and is most expressed by day 8, as expected. After successful establishment of spinal motor neuron differentiation, total RNA-seq datasets were used for circRNA identification and quantification.

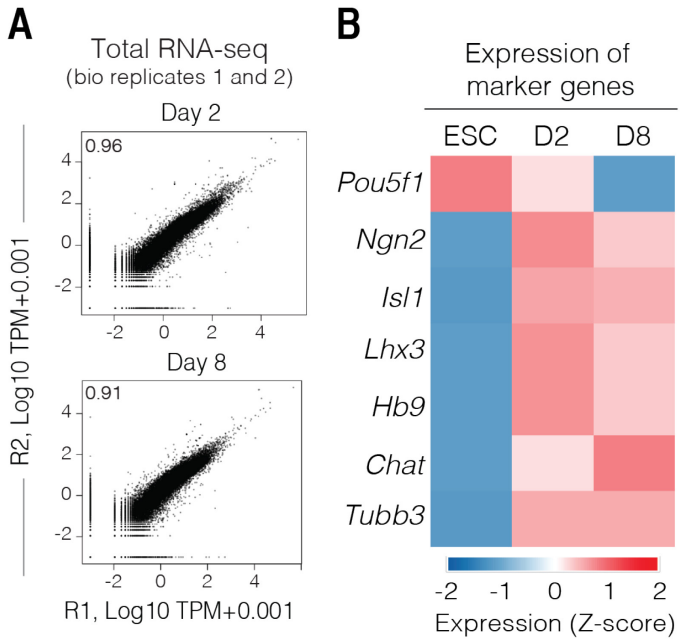

Figure 3.4 Marker gene expression during spinal motor neuron differentiation.
A) Pearson's correlation of gene expression for 2 biological replicates of induced EBs (day 2) and plated spinal motor neurons (day 8). **B)** Expression of marker genes after spinal motor neuron direct programming (D2 and D8) from total RNA-seq libraries. mESC-46C (ESC) gene expression represents mESCs state and is shown as comparison. Gene expression is depicted as Z-score.

3.6 CircRNAs are detected at all time-points and are most abundant in differentiated neurons

After careful validation of both neuronal systems, circRNAs were detected and quantified in two biological replicates of mESCs, and the dopaminergic and spinal motor neuron differentiation experiments. In total, 7773 circRNAs were detected, from which 1277 circRNAs were common in two biological replicates at any time-point (**Fig. 3.5A**). To determine whether the circRNAs detected in both replicates were more robust than the circRNAs detected in one biological replicate, I compared the number of back-spliced reads of common to uniquely detected circRNAs and found that circRNAs common in both biological replicates tend to have higher number of back-spliced reads and are therefore more reliably detected (**Fig. 3.5B**). This suggests that circRNAs uniquely detected are less expressed and at the limit of circRNA detection from the total RNA-seq datasets produced at the current depth of sequencing. Unless otherwise stated, the results presented in this thesis are based on circRNAs found in both biological replicates, but all analyses were also performed for each biological replicate separately and showed consistent results.

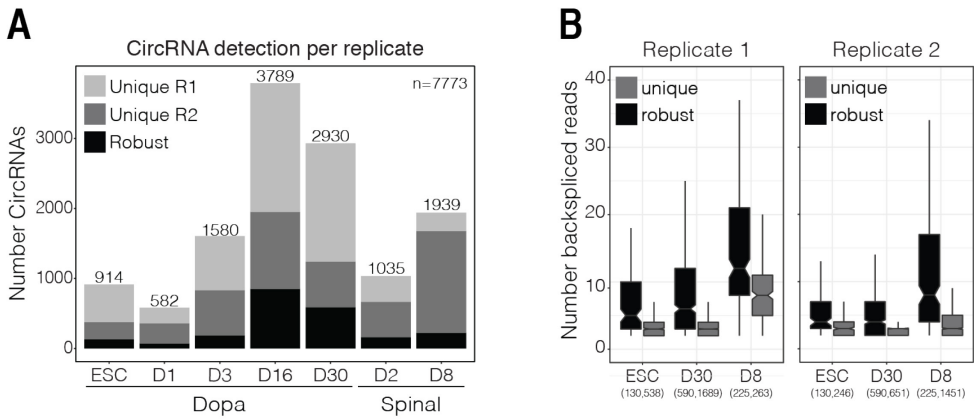

Figure 3.5 Comparison between circRNAs identified in both biological replicates.
A) Number of circRNAs detected uniquely (unique) or in both biological replicates (robust) in mESCs and during dopaminergic and spinal motor neuron differentiations. **B)** Number of back-spliced reads of robust and uniquely detected circRNAs for replicates 1 and 2, in mESC, dopaminergic neurons day 30 and spinal motor neurons day 8.

After assessing the quality of circRNA detection, I set out to understand which circRNAs are produced at each stage of dopaminergic and spinal motor neuron

differentiation. CircRNAs are detected at all time-points. However, the time-points with the highest number of circRNAs are days 16 and 30 of dopaminergic neuron differentiation (**Fig. 3.6A**, CircRNAs). The identified circRNAs are produced, in total, from 881 genes, with an amount at each time-point that follows a similar trend in the number of circRNAs (**Fig. 3.6A**, CircRNA-genes). Accordingly, most (~ 50 %) genes produce one circRNA, although some genes produce more than one (**Fig. 3.6B**). Expression of circRNAs from the same gene is more frequent on days 16 and 30, where the difference between the number of circRNAs and number of circRNA-producing genes is larger.

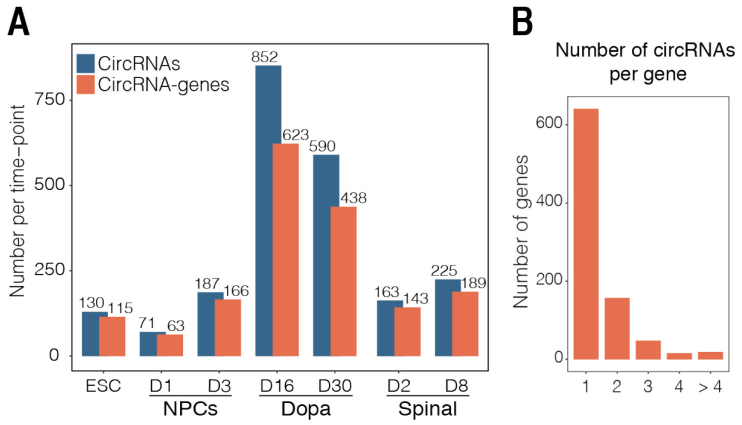

Figure 3.6 CircRNAs and genes producing these circRNAs are more abundant at later stages of neuronal differentiation.
A) Number of circRNAs and circRNA-producing genes at all time-points analysed. **B)** Number of circRNAs produced per gene for all genes producing circRNAs (n=881). Only circRNAs common in both biological replicates were considered for these analyses.

Next, I asked how the expression of individual circRNAs relates with gene expression. This can be studied by correlating the ratio of back-spliced and linearly-spliced reads (circ-to-linear ratio, a metric of exon-junction usage) with gene expression, which often shows a mild negative correlation (Rybak-Wolf et al. 2015). Gene expression was determined by calculating TPMs over the coding regions of genes from total RNA-seq datasets using RSEM software; similar results were obtained when considering mRNA-seq expression values (not shown). As expected, for all time-points considered in this study and in both biological replicates, circ-to-linear RNA ratio negatively correlates with gene expression (Pearson's correlation of -0.36 to -0.55, **Fig. 3.7)**, suggesting that

circRNA production may have a negative impact on the expression of the corresponding linear RNA species.

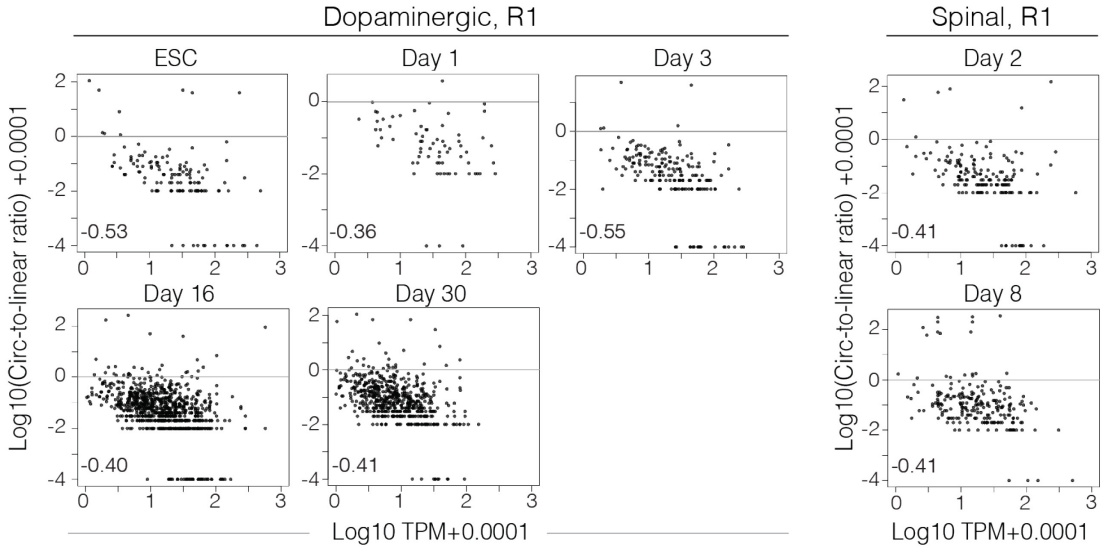

Figure 3.7 Relationship between back-splicing levels and expression of the linear transcript. Pearson's correlation of the ratio of the number of back-spliced reads and linearly spliced reads at exon junctions (circ-to-linear ratio) in relation to expression of the linear transcripts from the same gene (TPM) during dopaminergic and spinal motor neurons. Results are for circRNAs common to both replicates, and correlations are shown for replicate 1. Correlations were -0.53, -0.36, -0.55, -0.40, -0.41, -0.41 and -0.41, for ESC, and days 1, 3, 16, 30 of the dopaminergic and days 2 and 8 of the spinal motor neuron differentiations.

3.7 Expression features of genes producing circRNAs

3.7.1 Defining metrics to quantify circRNA expression per gene

The first step towards investigating the features of genes producing circRNAs was to identify an appropriate metrics to quantify circRNA expression per gene, considering that each gene may produce more than one circRNA and at varying levels of expression. To decide on an appropriate metric to express circRNA per gene, I started by investigating how the number of circRNAs and the abundance of back-spliced reads per gene relates with gene length. Exemplary data is presented here for mESCs, dopaminergic neurons day 16 and spinal motor neurons day 8; other time points showed similar correlations (data not shown).

When comparing the number of circRNAs per gene with host-gene length, there is no detectable correlation for all time-points tested (Pearson's -0.08 to 0.13, **Fig.**

3.8A). The number of back-spliced reads per gene also shows no correlation with host-gene length (Pearson's -0.01 to 0.11, **Fig. 3.8B**). Finally, when comparing the number of circRNAs detected with the number of back-spliced reads per gene, there is a mild positive correlation (Pearson's 0.08 to 0.33, **Fig. 3.8C**), suggesting that the more back-spliced reads are detected, the higher the number of circRNAs produced.

Based on these results, to quantify circRNA expression per gene, we normalized back-spliced reads to the number of circRNAs per gene, which was further normalized to all mapped reads in each dataset ($CircRP_{100}M$, **Fig. 3.8D**). To determine if the normalized back-spliced reads metric was representative of circRNA expression per gene, we correlated the number of back-spliced reads per gene with $CircRP_{100}M$. Indeed, both variables show a strong positive correlation for all time-points (Pearson's 0.73-0.95**, Fig. 3.8D**), suggesting that this metric is adequate to quantify circRNA expression per gene across all time-points of both neuronal differentiations.

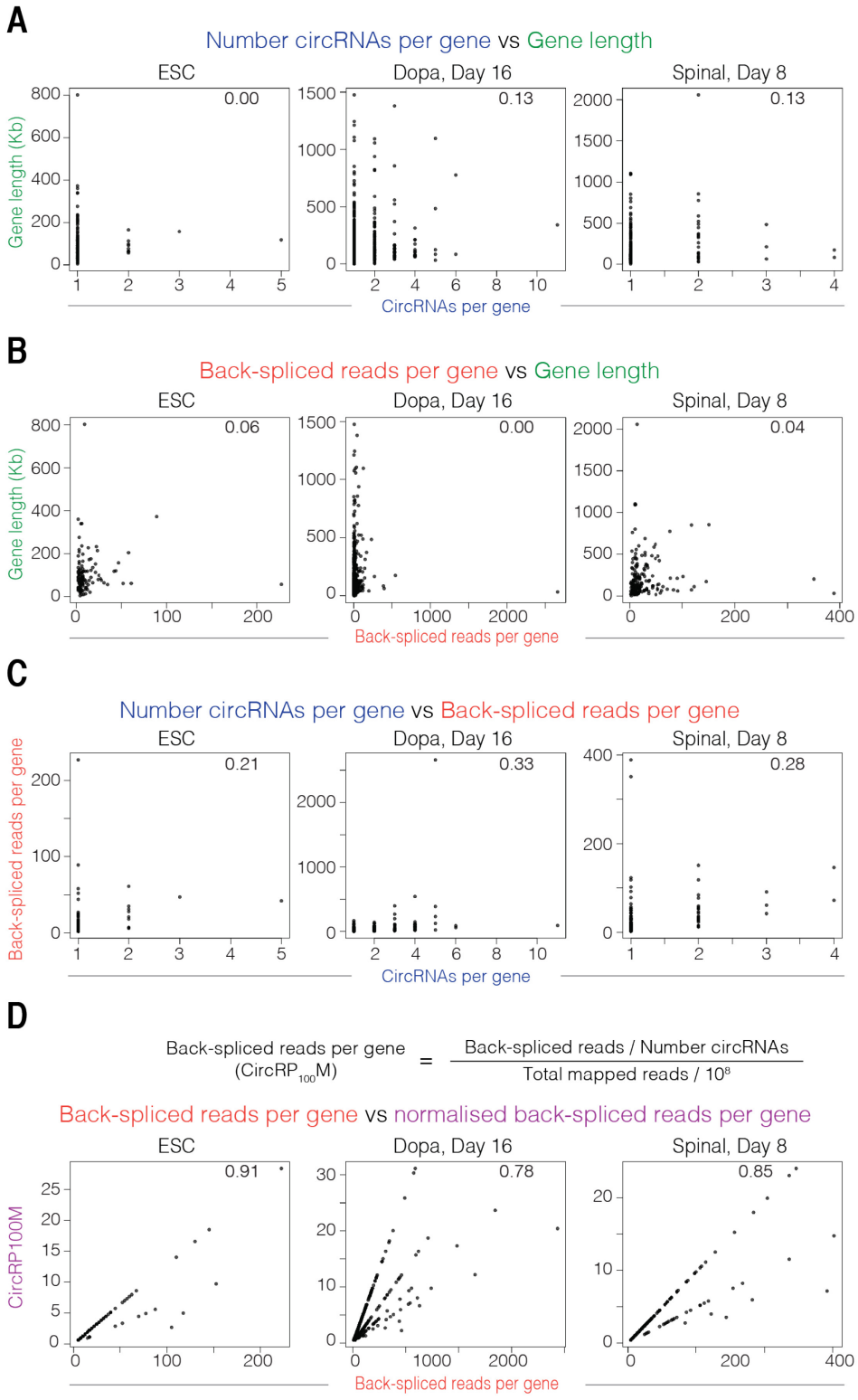

Figure 3.8 Selection of parameters for optimal quantification of back-spliced reads per gene. Pearson's correlation between: **A)** number of circRNAs per gene and gene length; **B)** number of back-spliced reads per gene and gene length; **C)** number of circRNAs per gene and number of back-spliced reads per gene; **D)** number of back-spliced reads per gene and normalised number of back-spliced reads per gene.

3.7.2 Genes producing circRNAs are expressed throughout differentiation, irrespective of circRNA expression

After defining a metric to quantify circRNA expression per gene, I set out to explore how circRNA expression relates with the RNA expression from genes producing circRNAs, calculated from total RNA-seq. I started by representing circRNA and gene expression from circRNA-producing genes in colored heatmaps, where genes were ranked according to circRNA expression at specific time-points. As is evidenced by **Fig. 3.9A**, most genes appear to express circRNAs at specific time-points, while some genes express circRNAs ubiquitously or in several time-points. Most genes produce circRNAs specifically at days 16, 30, or both, as previously stated. I next examined gene expression dynamics of circRNA-producing genes. I found that the majority of circRNA-producing genes is mostly expressed (TPM $\geq$ 1) throughout both differentiations, while few genes increase or decrease expression as differentiation progresses (**Fig. 3.9B**). Taken together, these results indicate that circRNA expression is a regulated process that does not depend exclusively on DNA sequence and/or gene expression.

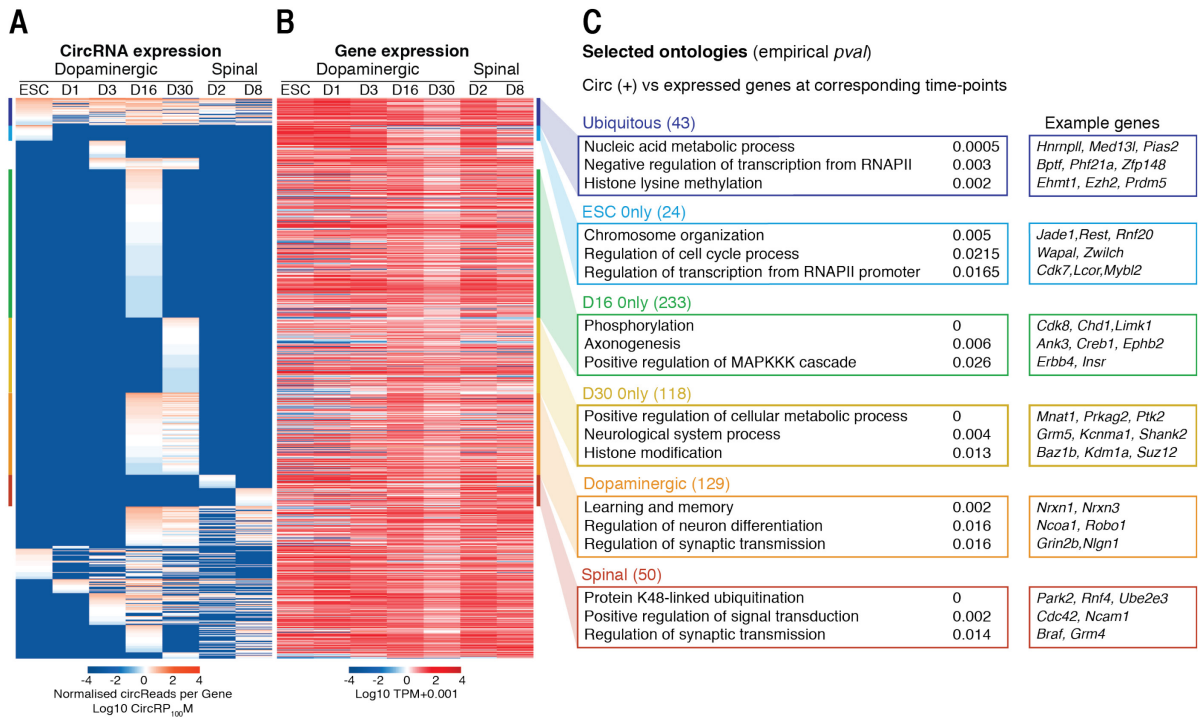

Figure 3.9 Characterization of genes producing circRNAs during dopaminergic and spinal motor neuron differentiations.
Heatmaps of **A)** circRNA expression per gene ($CircRP_{100}M$) and **B)** gene expression (TPM) throughout both neuronal differentiations. Total RNA-seq data was used and results are shown for biological replicate 1. Genes were ranked according to circRNA expression in the different cellular stages. **C)** GO terms enriched for genes producing circRNAs at specific stages of neuronal maturation. Representative GO terms were calculated using as background all genes expressed at corresponding time-points.

I next asked whether the genes that produce circRNAs at different stages of differentiation have specific biological functions. To address this, I performed GO analyses on groups of genes producing circRNAs at specific stages differentiation. These groups of genes were compared to all genes expressed at the corresponding time-point(s) (TPM >1). Reported enriched GO terms have at least 2 genes, a Z-score higher than 1.96, and permutation-derived p-values lower than 0.05 (see methods for further details). A small group of genes produces circRNAs at almost all time-points (Ubiquitous, n=43) and is enriched in GO terms important for basal cellular function, such as *"[GO:0090304] Nucleic acid metabolic process"* and *"[GO:0034968] Histone lysine methylation"*, including *Med13l* and *Ezh2* genes, which are important transcription regulators. Genes producing circRNAs only in mESCs (ESC only, n=24) have roles in transcription and cell cycle regulation, for example *Rest* and *Cdk7*, and are enriched in GO

terms *"[GO:0051276] Chromosome organization"*, *"[GO:0010564] Regulation of cell cycle process"* and *"[GO:0006357] Regulation of transcription from RNAPII promoter"*. Most genes produce circRNAs in immature and mature dopaminergic neurons and are enriched for terms related with neuronal maturation and function. Genes producing circRNAs only at day 16 (D16 only, n=233) are enriched for GO terms related with neuronal maturation and signaling, such as *"[GO:0016310] Phosphorylation"* and *"[GO:0007409] Axonogenesis"* (e.g. *Limk1, Ank3*, *Erbb4* genes), whereas genes producing circRNAs specifically at day 30 (D30 only, n=118) are enriched for GO terms related with metabolism and neuronal function, such as *"[GO:0031325] Positive regulation of cellular metabolic process"* and *"[GO:0050877] Neurological system process"* (e.g. *Ptk2*, *Grm5* and *Shank2* genes). Some genes produce circRNAs at both days 16 and 30 (Dopaminergic, n=129) and are enriched *for "[GO:0007611] Learning and memory"* and *"[GO:0050804] Regulation of synaptic transmission"* (e.g. *Nrx1/3*, *Grin2b* and *Nlgn1* genes). Finally, few genes produce circRNAs specifically in spinal motor neurons (Spinal n=50), which are also enriched for GOs related with neuronal function, such as *"[GO:0070936] Protein K48-linked ubiquitination"*, *"[GO:0009967] Positive regulation of signal transduction"* and *"[GO:0050804] Regulation of synaptic transmission*" (e.g. *Park2* and *Ncam1* genes). Overall these analyses show that although genes that produce circRNA are mostly expressed throughout differentiation, the expression of the circRNAs reflects differentiation stages at which the genes producing circRNAs are expected to have their cellular function, and therefore likely to be more highly expressed.

3.7.3 CircRNAs are produced when genes are most highly expressed

To further determine if circRNA-producing genes are highly expressed when producing circRNAs, I compared the expression of genes producing circRNAs at each time-point (circ+) with the expression of genes never producing circRNAs in any of the time-points considered in both differentiations (circ-). For the circ- gene group, I first included all genes that do not produce common or unique

circRNAs in any time-point biological replicates, and second, selected genes expressed with TPM ≥ 1 at each time-point. I also investigated the expression levels of all expressed genes at each time-point (Expressed, TPM ≥ 1). Strikingly, these analyses show that genes producing circRNAs at a given time-point are more highly expressed than the group of expressed and circ- genes at all time-points analyzed (**Fig. 3.10A**). Given that gene expression calculated from total RNA-seq data may be overestimated for genes with high levels of circRNA production, I repeated the analysis using published poly(A)-selected RNA-seq data that does not capture circRNAs, for all time-points of the dopaminergic neuron differentiation (Ferrai et al. 2017). Analyses of mRNA levels confirmed the results using total RNA-seq datasets, showing that genes that make circRNAs at a given time point are more highly expressed than expressed genes in the same time point, irrespectively of the potential of those genes for producing circRNAs in a different time point (**Fig. 3.10B**).

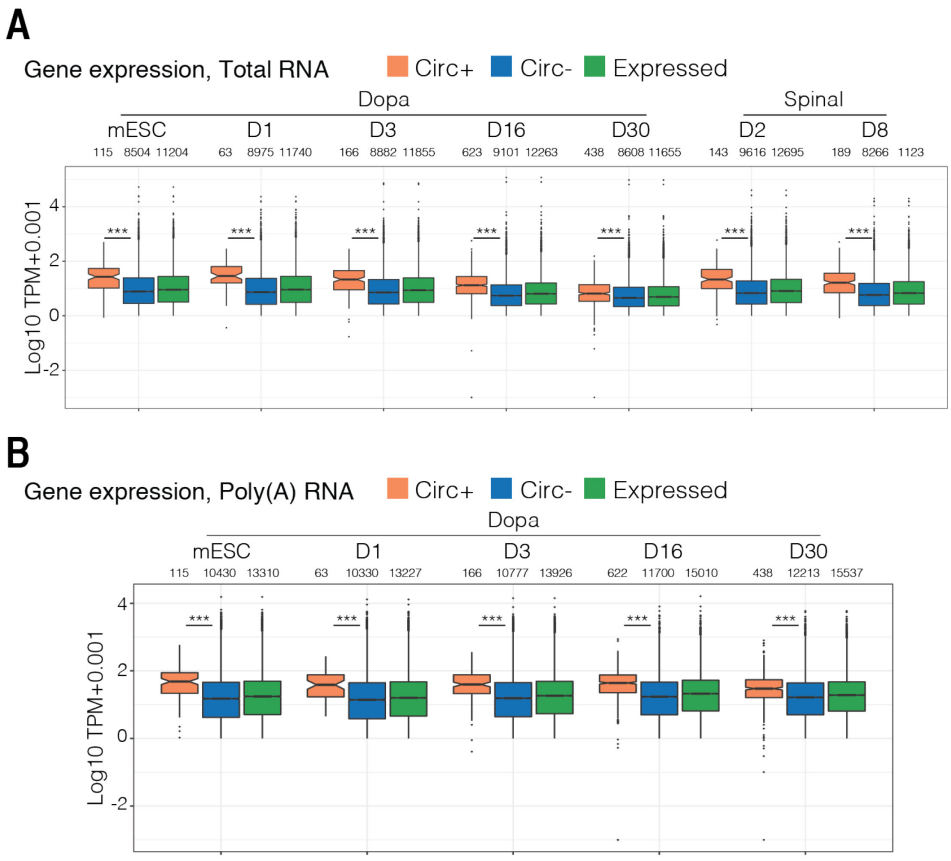

Figure 3.10 CircRNA-producing genes are highly expressed.
Boxplots show the comparison of gene expression (TPM) from **A)** total RNA-seq and **B)** poly(A) selected RNA-seq of genes producing circRNAs (Circ+), never producing circRNAs and expressed with TPM ≥ 1 (Circ-) and expressed with TPM ≥ 1 (Expressed) at all time-points of both neuronal differentiations. Numbers above boxplots represent number of genes in each gene group. n represents number of genes. Wilcoxon rank sum test, *** p-value < 0.001.

To further dissect the relationship between increased gene expression and circRNA production, I split all expressed genes (TPM ≥ 1) according to 5 quantiles based on gene expression, ranked from 1 (low) to 5 (high), and calculated the percentage of circRNA-producing genes in each quantile. Remarkably, for all time points analyzed, the proportion of circRNA-producing genes increases or stabilizes, suggesting that the more highly expressed a gene is in a given stage of differentiation the more likely it is to produce circRNAs. As an example, data is shown for mESC, dopaminergic neurons day 16 and 30, and spinal motor neurons day 8 (**Fig. 3.11**).

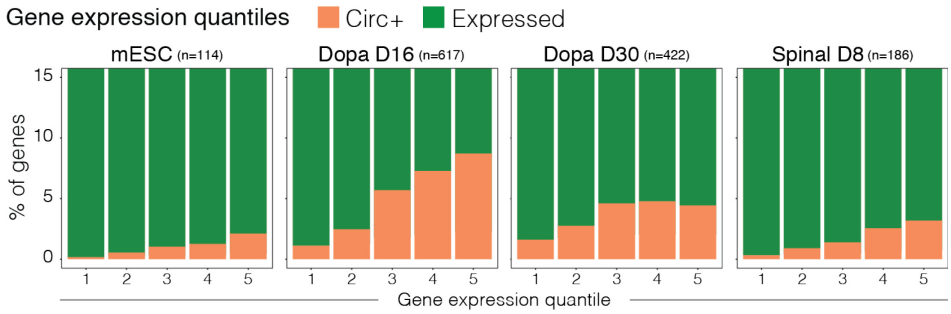

Figure 3.11 Percentage of circRNA-producing genes increases or stabilizes as gene expression increases.
Percentage of circRNA-producing genes per expression quantile of expressed genes at a given time-point. Genes were split in 5 quantiles ranging from 1 (low) to 5 (high). n represents number of genes.

3.8 Structural features of genes producing circRNAs

3.8.1 Genes producing circRNAs are very long and have many exons

Previous studies have shown that the structure of genes impacts circRNA formation; for example, long introns flanking the back-spliced exons or the presence of pairing repeats in introns on either side of the back-spliced exons promote circRNA production (Salzman et al. 2012; Ashwal-Fluss et al. 2014; Ivanov et al. 2015; Zhang et al. 2014). To investigate whether these features were common to the set of circRNAs detected in my datasets, I set out to explore the structural features of genes producing circRNAs during dopaminergic and spinal motor neuron differentiations. I started by comparing the gene length and number of exons of all circ+ and circ- genes. Genes expressed with TPM ≥ 1 at least once during either differentiation were used as control. For this analysis, I considered the transcript isoforms previously selected at each time-point, and calculated the mean transcript length and mean exon number for all groups of genes tested (see methods for further details). On average, I found that circ+ genes are much longer than circ- genes (~ 80 and 10 kb, respectively) and have more exons (~ 12 and 5 exons, respectively) (**Fig. 3.12A**).

To further dissect the relationship between gene length and circRNA production, I split all expressed genes according to 5 quantiles based on gene length ranked from 1 (low) to 5 (high), and calculated the percentage of circRNA-producing

genes in each quantile. For all time points analyzed, the proportion of circRNA-producing genes increases as gene length increases, suggesting that the longer the gene, the more likely it is that the gene produces circRNAs (**Fig. 3.12B**, data not shown for remaining datasets**)**. Altogether, these results suggest that genes producing circRNAs have a specific structure that may contribute to circRNA production, namely that they are longer and with more exons, and therefore may be more susceptible to imbalances in the recruitment of the splicing machinery to the chromatin.

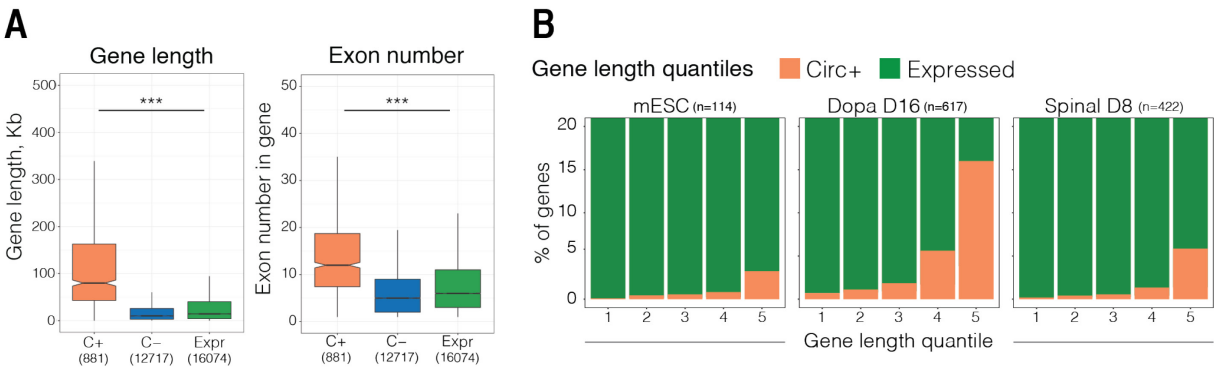

Figure 3.12 CircRNA-producing genes are longer and have many exons compared to genes not producing circRNAs.
A) Boxplots show comparison of gene length (kb) and exon number of genes producing circRNAs (C+), not producing circRNAs (C-) or expressed genes (Expr). Mean values of all transcript isoforms selected per time-point were used for this analysis. **B)** Percentage of circRNA-producing genes per length quantile of expressed genes at a given time-point. Genes were split in 5 quantiles ranging from 1 (low) to 5 (high). n represents number of genes. Wilcoxon rank sum test, *** p-value < 0.001.

3.8.2 CircRNAs are most often produced from the 5' end of genes and contain 1-5 exons

Next, I asked whether preferred exons are included in circRNAs. Previous studies showed that there is a tendency for circRNAs to be produced from the 5' end of genes (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012). To investigate whether this is the case in both differentiation systems used, I calculated the position of first back-spliced exon within the host gene. To ensure that transcript features were perfectly matched with circRNAs, I considered features of transcript isoforms attributed to circRNAs by the find_circ pipeline for this analysis. All non-exonic circRNAs (3 circRNAs) were excluded from this analysis. I

found that most circRNAs are made from the 5' end of genes, with a striking bias for starting from exon 2 (466 out of 1274, **Fig. 3.13A**). A small number of circRNAs start at exon 1, but after visual inspection on UCSC browser I noted that these circRNAs mostly result from genes only producing circRNAs (e.g. CDR1as) or that have upstream, misannotated transcription start sites. Performing the same analysis for individual time-points shows that circRNAs most often start at exon 2 irrespectively of the differentiation stage or cell type (**Fig. 3.13B**, data not shown). These results suggest that genes which produce circRNAs from exon 2 may have fewer exons, thus making it more likely that these genes produce circRNAs from exon 2. To address this, I compared the number of exons of transcripts producing circRNAs from exon 2 (E2) with transcripts producing circRNAs after exon 2 (>E2) or all transcripts producing circRNAs (All) (**Fig. 3.13C**). Results indicate that although transcripts producing circRNAs from exon 2 tend to have fewer exons, the difference is not sufficient to explain the bias towards increased circRNA production from exon 2. To further understand which exons are more often included in circRNAs, I calculated the number of exons included in circRNAs and found that most circRNAs have 1-5 exons, as expected (**Fig. 3.13D**) (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012). Accordingly, the position of the last back-spliced exon in the host-transcript most often ranges from exons 2 to 8 (**Fig. 3.13E**) (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012).

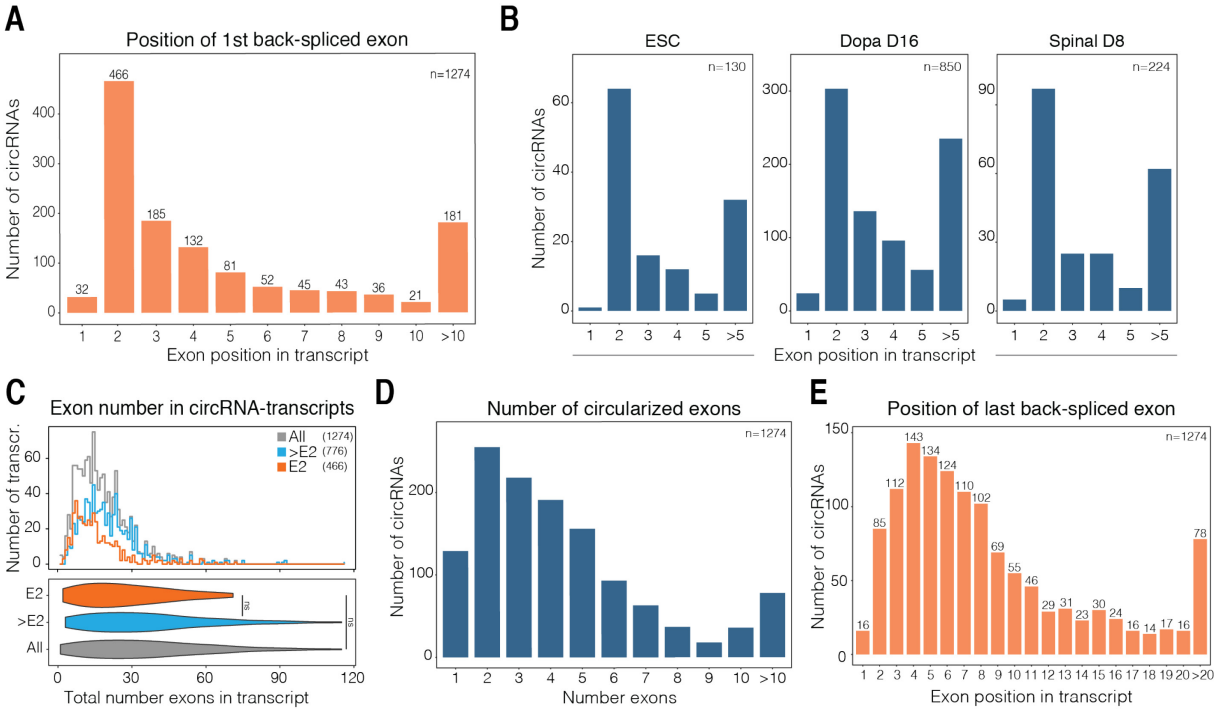

Figure 3.13 Features of back-spliced exons in circRNA-producing genes.
A, B) Position of the first back-spliced exon in the transcript, for A) all circRNAs or B) circRNAs produced at specific time-points. **C)** Total number of exons of transcripts producing circRNAs from exon 2 (E2), after exon 2 (>E2) or all circRNAs (All). **D)** Number of exons included in the circRNA, for all circRNAs. E) Position of the last back-spliced exon in the transcript, for all circRNAs. n represents number of circRNAs.

Given that circRNAs are most often produced from exon 2, I asked whether the first intron and exons of these genes have distinct features. I compared genes making circRNAs from exon 2 (E2) after exon 2 (>E2) or all circRNA-producing genes (All) with circ- (C-) genes at any given time-point; mean length of intron 1 and exons 1 and 2 were calculated across all time-points, as for the analyses shown in **Fig. 3.12** (see methods for further details). I found that the length of intron 1 of genes producing circRNAs from exon 2 is longer than circ- genes (19 and 1.6 kb, respectively, **Fig. 3.14A**), in agreement with previous reports that back-spliced exons being flanked by long introns (Salzman et al. 2012; Ashwal-Fluss et al. 2014; Ivanov et al. 2015; Zhang et al. 2014). However, genes producing circRNAs after exon 2 or all circRNA-producing genes show the same trend (9, 14 and 1.6 kb, respectively), suggesting that genes producing circRNAs tend to have a longer intron 1 irrespectively of the position of the first circRNA exon. This suggests that the longer first intron may be associated with deficient

linear splicing or deficient recruitment of the splicing machinery in circ+ producing genes.

To investigate in more detail gene structure aspects that may be associated with deficient linear splicing of the first intron, I considered the length of exon 1, which is not included in the circRNA and lies upstream of exon 2. I found that, on average, genes producing circRNAs from exon 2 tend have slightly shorter first exons than circ- (184 and 191 bp, respectively **Fig. 3.14B**). Finally, I found that genes producing circRNAs from exon 2 and all circRNA-producing genes tend to have a slightly longer exon 2 when compared to circ- genes (142 and 127 bp, **Fig. 3.14C**).

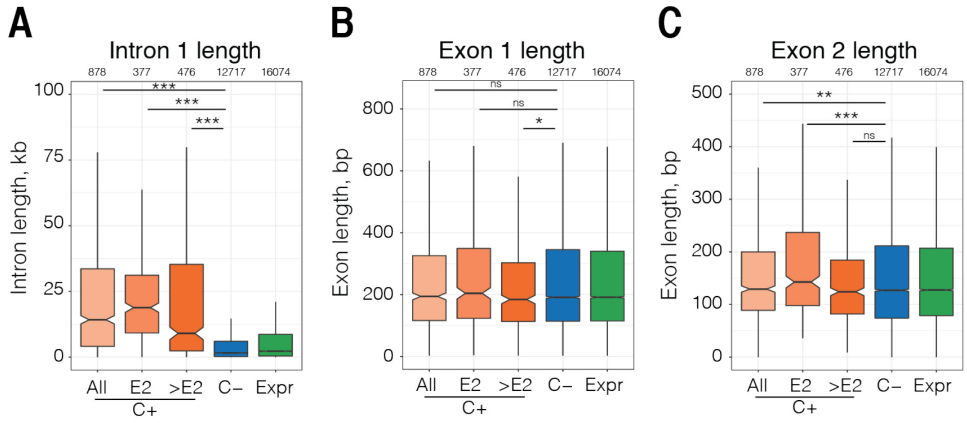

Figure 3.14 Comparison of intron and exon length between circRNA-producing genes and genes not producing circRNAs.
Boxplots show comparison of circRNA-producing genes (C+, All), genes producing circRNAs from exon 2 (C+, E2) or after exon 2 (C+, >E2) with genes not producing circRNAs (C-) or expressed genes (Expr) regarding **A)** intron 1 length (kb), **B)** exon 1 length (bp) and **C)** exon 2 length (bp). Wilcoxon rank sum test, * p-value < 0.05, *** p-value < 0.01, *** p-value < 0.001, ns, non-significant.

3.9 Discussion

3.9.1 Characterizing circRNA expression during neuronal maturation

CircRNA detection between biological replicates identified many common circRNAs and circRNA-expressing genes. A number of circRNAs detected were specific to each biological replicate, which may be due to their relatively low expression (**Fig. 3.5**); circRNAs found in both biological replicates were considered in subsequent analyses in this chapter.

CircRNAs were identified at all time-points of neuronal differentiation, but were highest at days 16 and 30 of dopaminergic neuron differentiation. When compared to dopaminergic neurons, circRNAs are much less abundant in spinal motor neurons and display numbers similar to mESCs or day 3 of the dopaminergic differentiation. This is likely due to the fact that the spinal motor neurons are not fully differentiated, and/or were pushed through an abnormally fast differentiation protocol. Alternatively, the lower abundance of circRNAs in the spinal motor neuronal samples may be explained by lower expression of genes that have the potential to produce circRNAs and/or decreased circRNA accumulation.

3.9.2 Genes producing circRNAs are highly expressed

To understand how circRNA expression from a given gene relates to its host-gene expression throughout differentiation, I quantified circRNA production per gene and compared it to gene expression. Similar to individual circRNAs, the extent of circRNA expression per gene is also very dynamic and cell-type specific. In contrast, most circRNA-producing genes are expressed at the mRNA level in all time-points investigated.

To understand whether the genes which produce circRNAs during dopaminergic and spinal motor neuron differentiations have specific biological functions, I performed GO enrichment analyses for groups of genes producing circRNAs at specific time-points. I found that genes producing circRNAs at specific time-points reflect the differentiation stage at which circRNAs are produced, raising further interest as to whether circRNAs have essential functions.

I found that circRNA expression coincides with high expression of the linear transcript from the same gene. This result may seem to contradict that circRNA-producing genes are most often expressed throughout differentiation; closer inspection showed that circRNA-producing genes show fluctuating gene expression during both differentiations with increased circRNA expression being associated with high expression of the linear transcripts. This finding is quite

surprising, as the current view in the field is that circRNA production competes with formation of the equivalent linear splicing, which leads to inferring that genes producing circRNAs would be expressed at low levels. The finding that genes producing circRNAs are highly expressed does not necessarily mean that circRNAs do not directly compete with the corresponding linear transcript. Instead, it suggests that abundant linear transcripts are produced from circRNA-producing genes, even though the ratio between back- and linear-spliced junctions tends to be negative. The finding that circRNAs are produced from highly expressed genes is further supported by the work of Zhang and colleagues, who showed that increased circRNA expression is associated with increased transcription from circRNA producing genes and with fast RNAPII elongation (Zhang et al. 2016). It is worth noting that a potential technical bias towards capturing circRNAs from highly expressed genes cannot be excluded, as genes expressed at lower levels could still produce circRNAs that are below the detection threshold in total RNA-seq datasets. Nevertheless, there are many highly expressed genes which do not produce circRNAs, suggesting that these results are biologically relevant. Finally, this finding raises the possibility that back-splicing events which lead to circRNA formation may be associated with excessive RNAPII loading and/or insufficient recruitment of the splicing machinery, especially upon removal of the first intron. Therefore, circRNA expression is likely to be a highly regulated process that depends on other regulatory factors beyond DNA sequence and gene expression.

3.9.3 Genes producing circRNAs are long, have many exons and most often produce circRNAs from exon 2

After exploring the expression dynamics of genes producing circRNAs, I set out to investigate their structural features. CircRNA-producing genes have a quite distinctive structure when compared to genes not producing circRNAs: these genes are very long and have many exons. Considering that the length of a gene mostly derives from the total length of its introns and that circRNAs are often flanked by very long introns (Salzman et al. 2012; Ashwal-Fluss et al. 2014;

Ivanov et al. 2015; Zhang et al. 2014), it is expected that genes producing circRNAs are very long. Another aspect that could contribute for increased length of circRNA-producing genes is that some of these genes are possibly neuronal and neuronal genes tend to be long (Zylka, Simon, and Philpot 2015). Additionally, circRNA-producing genes have many exons which may suggest that splicing of these genes is quite complex and requires extensive coordination between transcription and splicing machineries.

Next, I evaluated whether preferred exons were included in circRNAs. To address this, I calculated the number of exons included in circRNAs and the start and end position of circRNAs within the gene. Obtained results confirmed that circRNAs have most often 1-5 exons and that genes tend to produce circRNAs from the 5' end, with a striking bias for circRNAs starting at exon 2 (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012). This bias is independent of the differentiation stage or cell type and cannot be explained by transcript length alone. I further determined whether the first exons and intron of circRNA-producing genes displayed particular features by grouping genes according to circRNA production from exon 2 or after and compared to genes not producing circRNAs. All groups of genes producing circRNAs have a much longer intron 1. This suggests that intron 1 may be intrinsically more complex to splice, because the spliceosome subunits could have more difficulty identifying splice sites and forming the commitment complex. When looking at exon length, on average, genes producing circRNAs from exon 2 tend to have a slightly longer exon 2, while genes producing circRNAs after exon 2 have a slightly shorter exon 1. These slight differences might also contribute to which exons are included in circRNAs.

Taken together, results shown in this chapter point towards transcription and splicing dynamics underlying circRNA production. Genes producing circRNAs are highly expressed, which suggests high levels of transcription. Another remarkable finding was that a large number of circRNA-producing genes make circRNAs from exon 2, which points towards the first splicing reaction often being unfavored and giving rise to circRNAs. This suggests an imbalance between RNAPII loading and

spliceosome recruitment in genes producing circRNAs. In the next chapter I explore the mechanisms underlying transcription and spliceosome recruitment at genes producing circRNAs.

4 Promoter-based mechanisms of circRNA biogenesis

4.1 Research motivation and aims

The splicing of most introns is thought to happen co-transcriptionally (Ameur et al. 2011; Herzel, Straube, and Neugebauer 2018; Khodor et al. 2012; Nojima et al. 2018; Oesterreich et al. 2016; Tilgner et al. 2012) with splicing following a "first come, first served" model, where introns are spliced as soon as they emerge from the RNAPII exit channel. As is exemplified in **Fig. 4.1A**, at circ- genes, there is co-transcriptional recruitment of spliceosome subunits and formation of the commitment complex, followed by splicing. However, detection of exon-skipping and other alternative splicing events in many transcripts indicates that RNAPII can also elongate through several introns and exons before splicing is completed, thereby joining non-consecutive exons (Braunschweig et al. 2013; Drexler, Choquet, and Churchman 2019; Vuong, Black, and Zheng 2016; Drexler, Choquet, and Churchman 2020) The back-splicing reaction that leads to the biogenesis of circRNAs also requires that the intron upstream of the first back-spliced exon within any given circRNA is not immediately spliced.

We considered that at circ- genes, there is efficient co-transcriptional recruitment of spliceosome subunits and formation of the commitment complex, followed by efficient co-transcriptional splicing, especially of the first introns (**Fig. 4.1A**). However, when considering circ+ genes, we reason that transcription and splicing must be at least partially decoupled to allow circRNA formation (**Fig. 4.1B**). In more detail, two events are necessary for circRNA formation: firstly, the 1st back-spliced exon must be connected to the preceding intron so that it can accept the other back-spliced exons and form the circRNA; secondly, the intron preceding the 1st back-spliced exon cannot be spliced before RNAPII transcribes the remaining back-spliced exons. In this scenario, the first step necessary to

form circRNAs is that the intron before the 1st back-spliced exon is not immediately spliced after being transcribed.

Evidence suggests that circ+ genes seem to have altered transcription-splicing dynamics. Some genes display increased circRNA production upon knockdown of splicing components (Liang et al. 2017) and circ+ genes tend to be transcribed by a fast elongating RNAPII (Ashwal-Fluss et al. 2014; Zhang et al. 2016). Together, these results suggest that the spliceosome may not be efficiently recruited to these genes.

In the previous chapter, I showed across seven different stages of differentiation to mature neuronal lineages, that circ+ genes are expressed at high levels. I also showed that genes produce circRNAs mostly from exon 2, involved in the first splicing reaction, for which a defect in recruitment of the splicing machinery may possibly occur due to increased speed of the RNAPII as it leaves the promoter. However, how transcription-splicing dynamics contribute to circRNA formation remains largely unexplored. Is spliceosome recruitment at the first exon-intron junction altered at circ+ genes? Or is there insufficient modification of RNAPII at initiation, or altered exit from promoter-proximal pausing? RNAPII post-translational modifications are essential for the proper co-transcriptional recruitment of the spliceosome complex and it remains unexplored whether RNAPII modifications and transcription regulation contribute to circRNA formation.

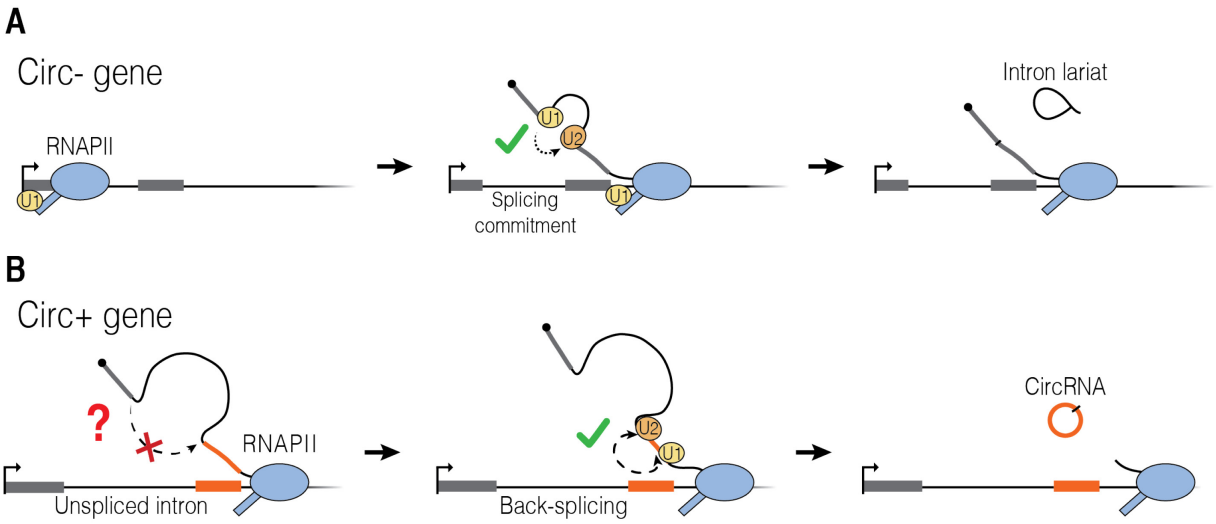

Figure 4.1 Proposed steps of circRNA formation.
A) Illustration of co-transcriptional splicing at genes not producing circRNAs (circ-). U1 snRNP is co-transcriptionally recruited and commitment complex is formed as RNA exits RNAPII followed by splicing and release of intron lariat. **B)** At genes producing circRNAs (circ+), the intron preceding the back-spliced exons is not immediately spliced so that the splicing acceptor is free to later join the back-spliced exon. As such, the back-splicing would be favoured and lead to circRNAs production.

To answer these questions, I mapped and analyzed the occupancy on chromatin of the spliceosome, RNAPII post-translational modifications and transcription modulators in mESCs, dopaminergic neurons (days 16 and 30) and spinal motor neurons (days 2 and 8). To this end, I generated new datasets and re-analyzed published datasets. I show that circ+ genes have decreased spliceosome recruitment at the promoter and first exon-intron junction compared with circ- genes. The lower levels of spliceosome recruitment found at the first exon-intron junction is not due to defective RNAPII recruitment, but coincides with lower RNAPII-S5p and -S7p enrichment. Furthermore, circ+ genes are depleted of promoter-proximal pausing modulators, suggesting that RNAPII is quickly released from the promoter into productive elongation without the appropriate CTD modifications and with impaired spliceosome recruitment. Analyses of circRNA formation in different cell types, supports similar promoter-based mechanisms underlying circRNA formation in dopaminergic and spinal motor neurons. Finally, to further test the proposed model that decreased promoter-proximal pausing may modulate to circRNA formation, I knocked-down (KD) NELF complex to trigger RNAPII release from the promoter and increase circRNA

production. As predicted by the proposed model, NELF KD increased the number of circRNAs detected.

4.2 Contribution disclosure

From Pombo group: Alexander Kukalev contributed to the production of ChIP-seq libraries (S5p, Dig and Mock in mESCs) and taught me ChIP-seq quality controls, as well as designed the interference RNAs for NELF knockdown experiments. Izabela Harabula provided chromatin samples for ChIP of U1C and NELF-E in dopaminergic neurons day 16. Tiago Rito devised the strategy for gene filtering analyses to define circ+ and circ- genes, guided me through ChIP-seq analyses and provided the script for calculating the coverage of ChIP-seq enrichment. Christoph Thieme, Warren Winick-Ng and Dominik Szabo advised on the statistical analyses of NELF knockdown experiments.

From Nikolaus Rajewsky group, MDC: Petar Glažar performed circRNA identification together with a list with matched linear transcripts and performed corresponding quality controls for the NELF knockdown experiments.

From Esteban Mazzoni group, NYU: Silvia Velasco provided the protocol for ChIP in embryoid bodies which I later adapted for RNAPII ChIP.

4.3 Notes to the reader

Analyses for biological replicates 1 and 2 were performed in parallel and results were consistent between biological replicates. Results shown correspond to biological replicate 1, unless otherwise specified. Analyses shown in this chapter considered only circRNAs robustly detected in two biological replicates, except in NELF knockdown experiments.

4.4 Mapping spliceosome and RNAPII modifications on chromatin

To understand if co-transcriptional splicing dynamics are altered at circRNA-producing genes, I started by mapping the occupancy of the spliceosome and RNAPII post-translational modifications on chromatin in mESCs. I chose to map U1C protein, a subunit of the U1 snRNP complex, because not only U1 snRNP is the first spliceosome subunit recruited to splice sites, thus playing a key role in splice site definition and commitment complex formation, but also interacts with RNAPII S5p (Harlen et al. 2016; Nojima et al. 2015; Nojima et al. 2018). As expected, U1C is enriched at the TSS of expressed genes, being more highly enriched at highly expressed genes (top 20% expressed genes) when compared to lowly expressed genes (bottom 20%) or genes that are not expressed (Not Expr) (**Fig. 4.2A**). Consistent with U1 snRNP's role in 5'ss definition, U1C also shows some enrichment at exons and 5'ss (**Fig. 4.2B**).

To identify the genomic regions that are enriched for U1C above background, I performed peak finding with the BCP algorithm (see methods for further details), (Ferrai et al. 2017), and classified all the promoters (windows of 4kb centered on the TSS) taking advantage of a procedure established in our group by Dr. Elena Torlai-Triglia and Dr. Tiago Rito, (see methods for further details), (Ferrai et al. 2017). The distribution of ChIP-seq read counts often shows a bimodal distribution that separates signal from noise. In U1C data, the distribution of read counts corresponding to signal and noise correspond to a tall peak of higher values together with a shoulder towards lower levels of read abundance (**Fig. 4.2C**). The lack of a clear bimodal separation is frequent in this type of analyses, and can result from a small differential between the number of reads found in a positive U1C window compared with the negative windows, which may be due to difficulties recovering U1C, which does not bind directly to chromatin but, instead to nascent RNA. As a further test of specificity, I compared the abundance of reads in the U1C positive windows with shuffled BCP windows, and found that they give lower (noise) read counts when compared to U1C enrichment (**Fig. 4.2D**).

Altogether, these analyses show that U1C ChIP-seq is suitable for further exploration.

I next set out to investigate whether the production of circRNAs from specific genes in mESCs reflected altered RNAPII post-translational modifications. Previous work in our lab had produced RNAPII S5p and S7p datasets in mESC (Brookes et al. 2012; Ferrai et al. 2017). However, since the protocol for library preparation of U1C ChIP-seq was slightly different, I produced fully matched ChIP-seq datasets for RNAPII S5p, S7p and S2p. Visual inspection of ChIP-seq tracks on the UCSC browser showed that these marks had the occupancy expected in mESC data from published literature (e.g. (Brookes et al. 2012; Ferrai et al. 2017); not shown).

Next, I classified gene promoters according to whether they are occupied by RNAPII-S5p and S7p; S2p is not enriched at promoters. To identify positively enriched promoters in S5p and S7p, I determined the density of read counts at TSS and their distribution shows the typical bimodal pattern, indicating a good signal-to-noise ratio (**Fig. 4.2E**). Detection of positive peaks with BCP at promoters shows that positive windows overlap extensively between matched datasets collected previously (Ferrai et al. 2017) (11765 out of 13444 peaks for S5p, and 9856 out of 11221 peaks for S7p). The enrichment level of positive windows common in matched datasets correlates very well (Spearman's correlation, 0.75 and 0.88, respectively). Altogether, analyses show that produced datasets are highly comparable to published data (**Fig. 4.2E**).

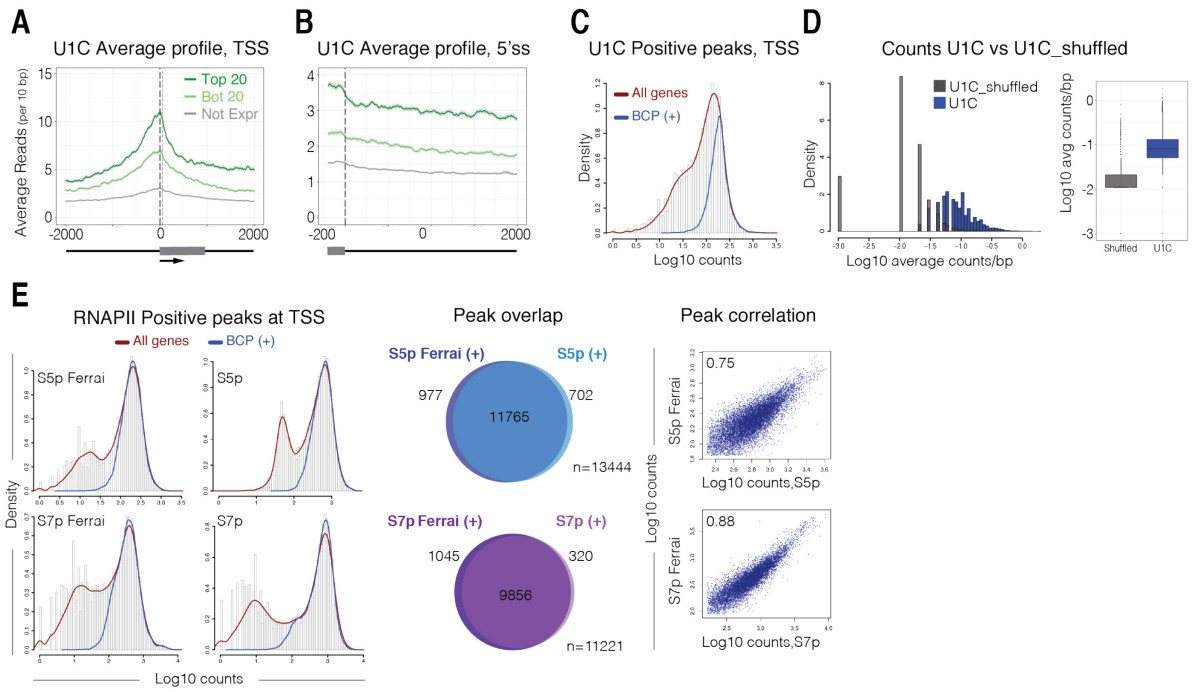

Figure 4.2 Validation of U1C, RNAPII S5p and S7p ChIP-seq in mESCs.
A) Average distribution of U1C in 4 kb windows centered around the TSS of top 20%, bottom 20% and genes that are not expressed (TPM<1). **B)** Average distribution of U1C in windows of -200 bp to 2kb around 5' splice sites. All introns shorter than 2000 bp were not considered in this analysis. **C)** Density of reads of U1C dataset in promoter windows of all genes compared with genes classified as positive by BCP. **D)** Comparison of read density and counts between BCP-positive U1C peaks at TSS and shuffled peaks. Boxplot shows higher read count in U1C BCP-positive peaks when compared to shuffled peaks. **E)** Comparison between published and newly produced datasets for RNAPII S5p and S7p. Left panel: Density of reads of RNAPII S5p and S7p dataset in promoter windows of all genes classified compared with genes classified as positive in BCP. Middle panel: overlap of BCP positive peaks at TSS between matching datasets. Right panel: Spearman correlation of read counts at TSS between matching datasets.

I next set out to explore the patterns of chromatin occupancy of U1 snRNP, total RNAPII detected with N20 antibody (published by the Young lab, see methods), and RNAPII modifications at circRNA-producing genes in mESCs. Results from the previous chapter showed that most genes produce circRNAs from exon 2, which indicates that the first splicing reaction is most often missed and suggests that spliceosome recruitment may be altered at the beginning of genes. I started by inspecting circ+ and circ- genes on UCSC genome browser and focused on the TSS of genes with comparable mRNA expression levels present in the third and fourth quantiles of expressed genes. Circ+ genes appeared to show less enrichment of U1 snRNP and RNAPII S5p and S7p at the TSS, whereas RNAPII S2p and total RNAPII were more similar, suggesting that U1 snRNP recruitment and RNAPII

patterns may be altered at circ+ genes. An example is shown on **Fig. 4.3** for *Gtf3c3* (circ-) and *Rabep1* (circ+) genes.

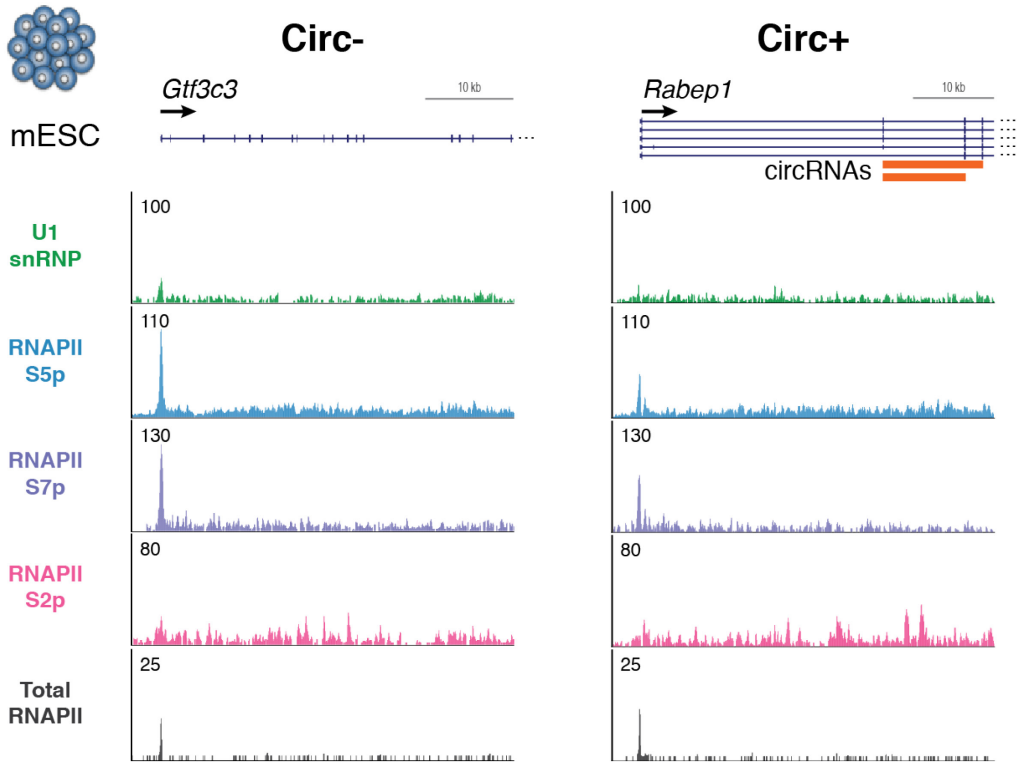

Figure 4.3 Single gene profiles to illustrate the occupancy of U1 snRNP and RNAPII in circ- and circ+ genes.
Examples to highlight the differences in enrichment of U1 snRNP and RNAPII modifications between circ- gene *Gtf3c3* and circ+ gene *Rabep1* in mESCs. These genes have comparable mRNA expression levels (TPM = 26.37 and 21.65, respectively) and belong to the third quantile of expressed genes.

4.5 Approach to study occupancy of the spliceosome and RNAPII modifications at circRNA-producing genes

To understand the dynamics between spliceosome recruitment and RNAPII modifications at circRNA-producing genes in a systematic and quantitative manner, we took an approach which compares all circRNA-producing genes at a given time-point (circ+) with genes never producing circRNAs at any time-point of both differentiations (circ-). Circ+ genes are more highly expressed than circ- and the amount of RNAPII modifications at promoters was shown to be most predictive of gene expression levels (Dias et al. 2015). To identify a group of circ- genes with comparable mRNA expression to circ+ genes, I explored several

thresholds for gene expression and found an optimal range between 25% and 80% of the circ+ gene expression (data not shown). For keeping track of how expression, U1C and RNAPII occupancy is related to circRNA production, I also included two additional group of genes: expressed (TPM ≥ 1) and not expressed (TPM < 1). Additional filters were used: all genes in each group have at least 3 exons, as this is the minimum number of exons required to produce circRNAs from protein-coding genes (**Fig. 4.4**, left panel). Since circRNAs most often result from the first splicing reaction, I focused my analyses on the TSS and exon 1 – intron 1 border (E1-I1), with windows spanning 1 kb (**Fig. 4.4**, right panel). For these regions, I calculated the average profiles (dark blue line) and enrichment (light blue shade) of spliceosome and RNAPII modifications for the above-mentioned gene groups (see methods for further details). Other explorations were carried out, including at intron1-exon2 junctions, but not included here since they did not show distinct patterns.

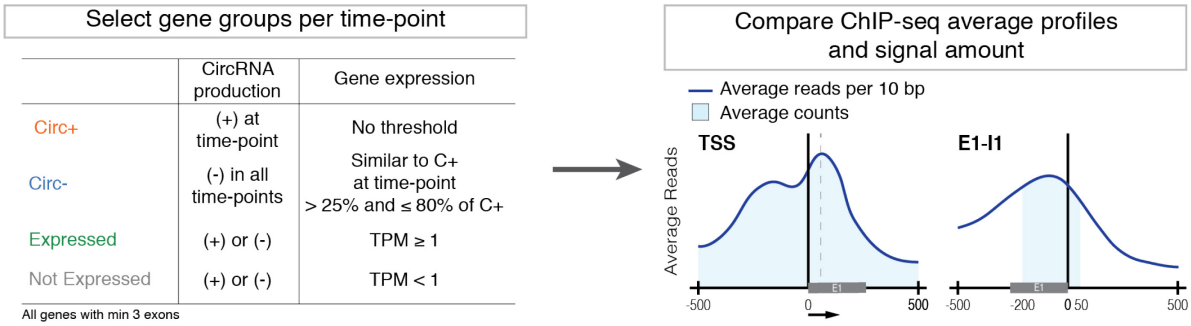

Figure 4.4 Schematic overview of the approach to study enrichment of proteins on chromatin of circRNA-producing genes.
Left panel: description of criteria for the selection of different gene groups. Circ+, all genes producing circRNAs at a given time-point, with minimum of 3 exons. Circ-, all genes never producing circRNAs during differentiation, with matched expression of circ+ genes of the corresponding time-point and with at least 3 exons. Expressed: all genes with TPM ≥ 1 at a given time-point. Not Expressed: all genes with TPM < 1 at a given time-point. **Right panel:** after selecting gene groups, compare average enrichment and read counts of several ChIP-seq datasets in 1kb windows centered at the TSS or exon 1-intron 1 border (E1-I1). The light blue area represents the window used to determine read counts at the TSS (1kb window) and E1-I1 border (-200 to 50 bp).

4.6 Promoter-based regulation of circRNA biogenesis in mESCs

4.6.1 The spliceosome is depleted at the 5' end of circRNA-producing genes

To investigate the patterns of occupancy and the enrichment of the spliceosome and RNAPII modifications in mESCs, I compared gene expression levels, and the length of Intron1, Exon1 and Exon2 of circ+, circ-, expressed and not expressed gene groups defined according to the above-mentioned parameters. I found that circ+ and circ- gene groups show comparable gene expression levels. In contrast, circ+ genes have longer intron 1 than circ- genes (median 18 kb compared with 1kb), and slightly longer exon 1 (250 bp vs 150 bp) and exon 2 (125 bp vs 110 bp) (**Fig. 4.5**). Expressed genes are shown for comparison.

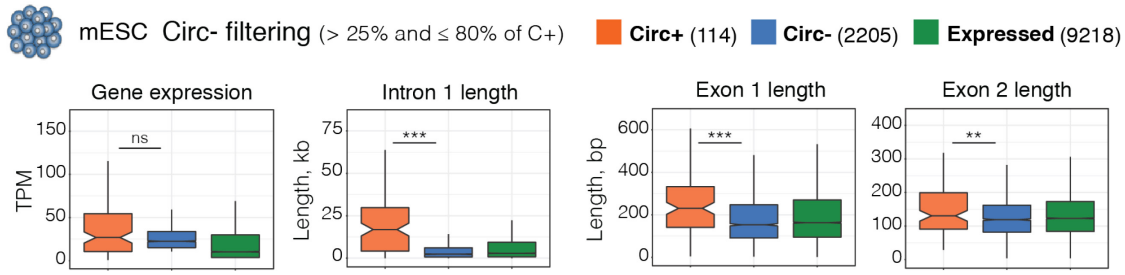

Figure 4.5 Circ- gene filtering to match circ+ expression in mESCs.
Boxplots showing gene expression levels, intron 1 and exons 1 and 2 length for circ+, circ- and expressed genes in mESCs. Circ- genes were selected with a threshold from 25% to 80% of circ+ expression. Significance determined with Wilcoxon rank sum test; *** p-value < 0.001.

I next compared the occupancy of U1 snRNP (U1C subunit) on chromatin at the TSS and E1-I1 border of circ+ and circ- genes, and found that U1 snRNP occupancy is depleted at circ+ genes (**Fig. 4.6**). This is visible in both average profiles and boxplots of enrichment levels at the TSS and E1-I1 border. These results are consistent with decreased U1 snRNP recruitment and altered recognition of the first 5'ss at circ+ genes.

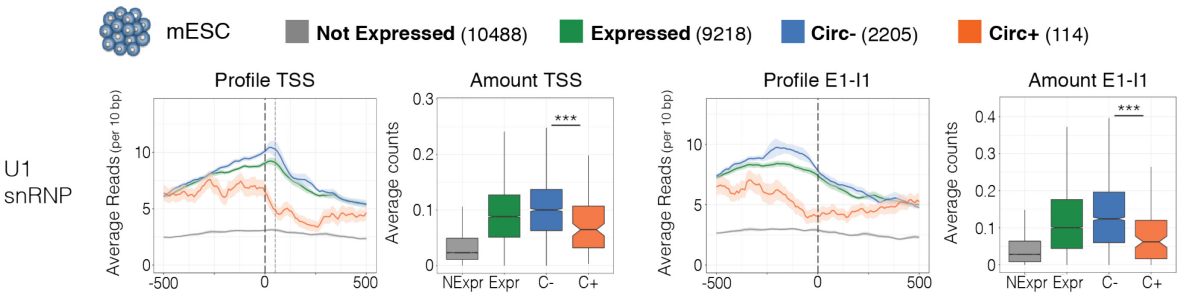

Figure 4.6 Enrichment of U1 snRNP at the TSS and E1-I1 border in mESCs.
Average distribution of U1 snRNP in 1kb windows at TSS and E1-I1 border. Boxplots show amount of U1 snRNP at these locations. Significance determined with Wilcoxon rank sum test; *** p-value < 0.001.

4.6.2 RNAPII S5p and S7p are depleted at circRNA-producing genes, while S2p is unchanged

RNAPII modifications play an essential role in the co-transcriptional recruitment of protein complexes that process nascent RNA and that U1 snRNP was shown to directly bind RNAPII S5p as it transcribes (Harlen et al. 2016; Nojima et al. 2015; Nojima et al. 2018). To study whether spliceosome depleted at the promoter regions of circ+ genes is due to altered RNAPII occupancy or modification, I examined the enrichment of RNAPII modifications at the TSS and E1-I1 border of circ+ and circ- genes. I found that RNAPII S5p is depleted at the TSS and, more profoundly, at the E1-I1 border of circ+ genes when compared to circ-, which correlates with altered U1 snRNP recruitment at the promoter of circ+ genes (**Fig. 4.7A**). Remarkably, RNAPII S7p is also depleted at the TSS and E1-I1 border of circ+ genes (**Fig. 4.7B**). Given that S7p plays an important role in the transition between transcription initiation and elongation, these results may hint on promoter-proximal pausing being altered at circ+ genes. In contrast, RNAPII S2p, which is associated with productive elongation and contributes to spliceosome recruitment at later stages of the transcription cycle, is only slightly decreased at the TSS and E1-I1 border or unchanged at the TES, where its enrichment is most prominent (**Fig. 4.7C**).

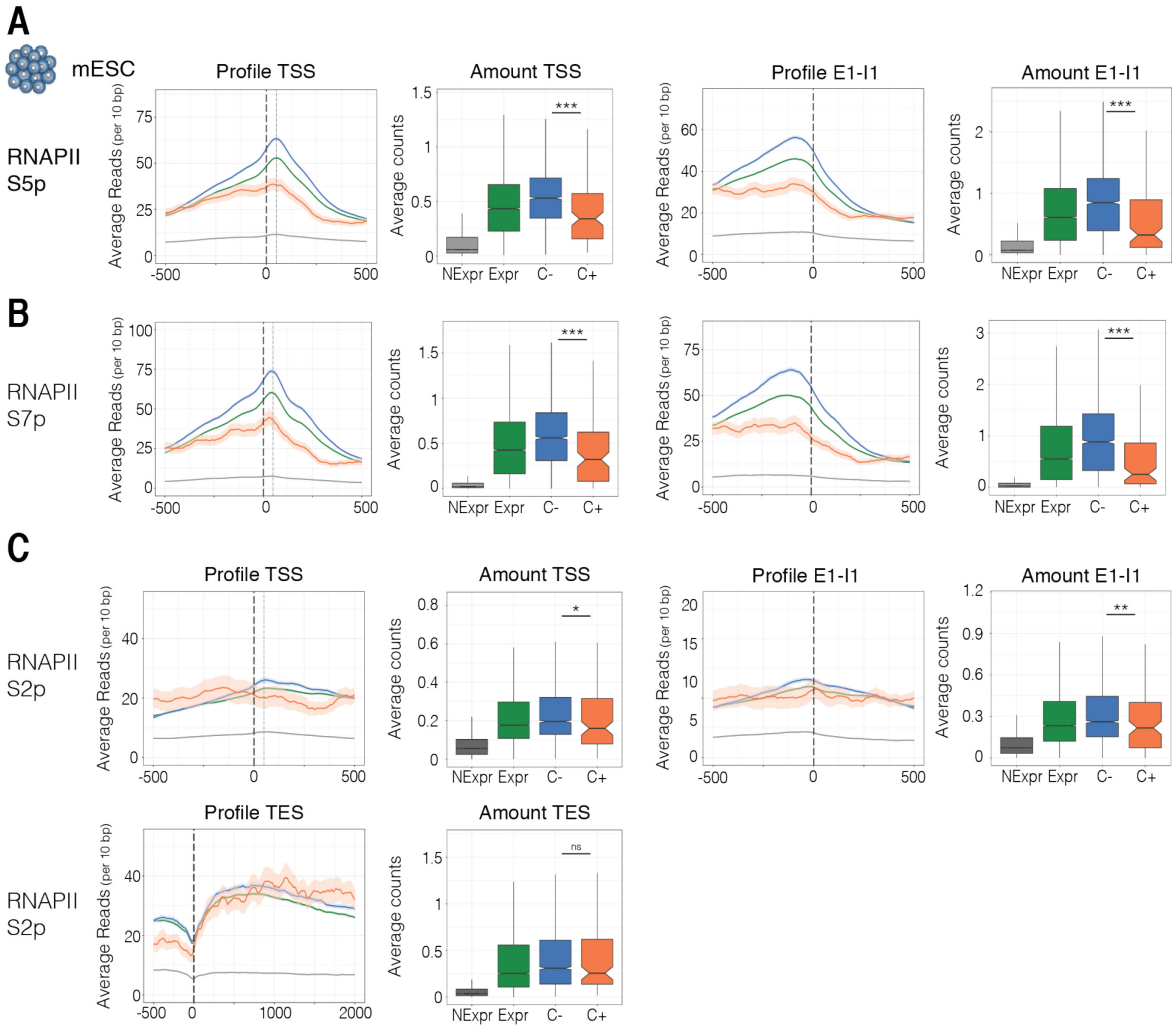

Figure 4.7 Enrichment of RNAPII post-translational modifications at the TSS and E1-I1 border in mESCs.
Average distribution of **A)** RNAPII S5p, **B)** S7p and **C)** S2p in 1kb windows centered at TSS and E1-I1 border. Boxplots show amount of RNAPII post-translational modifications at these locations. Significance determined with Wilcoxon rank sum test; * p-value < 0.05, ** p-value < 0.01; *** p-value < 0.001; ns – not significant.

Negative controls for RNAPII ChIP-seq using the unspecific antibody Digoxigenin (Dig) or beads incubated in chromatin extract (Mock) show a slight decrease at the TSS of circ+ genes. Nevertheless, this difference is very small in comparison to specific IPs and therefore negligible (**Fig. 4.8**).

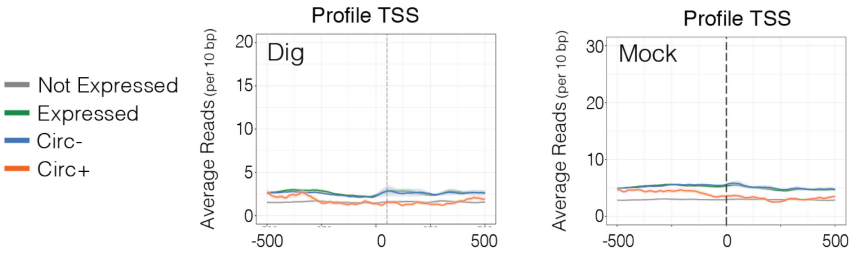

Figure 4.8 Enrichment of negative controls at the TSS and E1-I1 border in mESCs.
Average distribution of ChIP with an anti-digoxigenin antibody (Dig) or in the absence of antibody (Mock) in 1kb windows centered at TSS show no enrichment.

To understand if RNAPII S5p and S7p depletion at circ+ genes could be explained by decreased RNAPII loading, I explored a published total RNAPII ChIP-seq dataset (N20 antibody, see methods for details) in mESC. I found that total RNAPII is slightly decreased at both TSS and E1-I1 border of circ+ genes (**Fig. 4.9**). However, total RNAPII depletion is not as striking as the depletion of S5p and S7p, suggesting that the depletion of S5p and S7p cannot be explained by decreased RNAPII loading alone, but also suggest decreased modification. Altogether, these results raise the possibility that circ+ genes display altered dynamics between transcription and splicing in early stages of the transcription cycle close to the promoter regions, at the time when spliceosome complexes are recruited by RNAPII to chromatin to process nascent RNA co-transcriptionally.

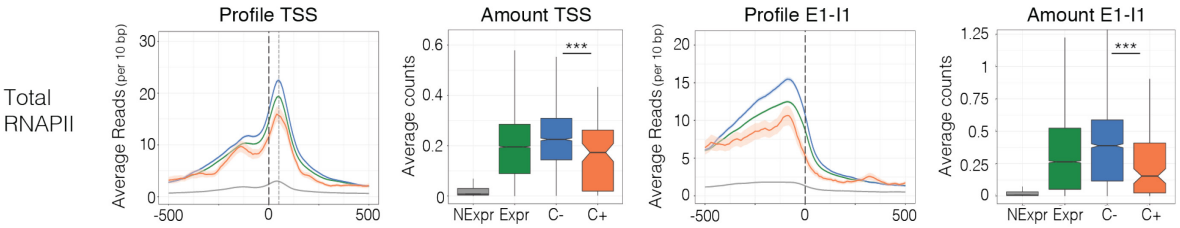

Figure 4.9 Enrichment of total RNAPII at the TSS and E1-I1 border in mESCs.
Average distribution of total RNAPII in 1kb windows centered at TSS and E1-I1 border. Boxplots show enrichment at these locations. Significance determined with Wilcoxon rank sum test; *** p-value < 0.001.

4.6.3 Factors that regulate promoter-proximal pausing are depleted at circRNA-producing genes

Given that RNAPII S5p and S7p are depleted at circ+ genes and that total RNAPII levels cannot fully explain these results, we reasoned that one or several steps in the transcription cycle could be altered. For example, RNAPII might be less recruited to circ+ genes, less phosphorylated on S5 and S7 upon transcription initiation, or be quickly released from initiation into elongation. To address this, I took advantage of several published ChIP-seq datasets in mESCs (see Table 2.7 for details) and determined the occupancy on chromatin of general transcription factors (TBP and TAF1), RNAPII transcription initiation factors (CDK7 and 8) and promoter-proximal pausing modulators (NELF-A and CDK9).

In addition to the use of the published datasets, I also mapped NELF-E by ChIP in mESCs, which is an additional subunit of the NELF complex. Before exploring the occupancy of the published and new ChIP-seq datasets, I started by mapping and confirmed the quality of the new NELF-E dataset. First, I checked whether NELF-E is detected at the promoter regions of active genes as expected, by comparing average ChIP-seq profiles of NELF-E at the most and least expressed genes (**Fig. 4.10A**). The average profile of NELF-E shows enrichment at the TSS of the top 20% most expressed genes (TPM $\geq$1), intermediate enrichment at the 20% least expressed genes, and no detectable enrichment at not expressed genes (TPM$<$1). Comparisons of the NELF-E ChIP-seq dataset produced in the present study with the published NELF-A dataset shows that the new dataset has improved signal-to-noise ratio (**Fig. 4.10B**).

Next, I calculated regions enriched for NELF-E and NELF-A using BCP and Mock as control ChIP-seq dataset. After classification of promoters according to whether they overlap with NELF-E and/or NELF-A peaks, I found that they mark the TSSs of the same genes, with extensive overlap (10555 out of 11857 positive peaks) and whose enrichment level correlates well (Spearman's correlation 0.83) (**Fig. 4.10C)**. Finally, NELF-E positive windows show higher read counts when

compared to shuffled positive windows, indicating that NELF-E enrichment is specific (**Fig. 4.10D**).

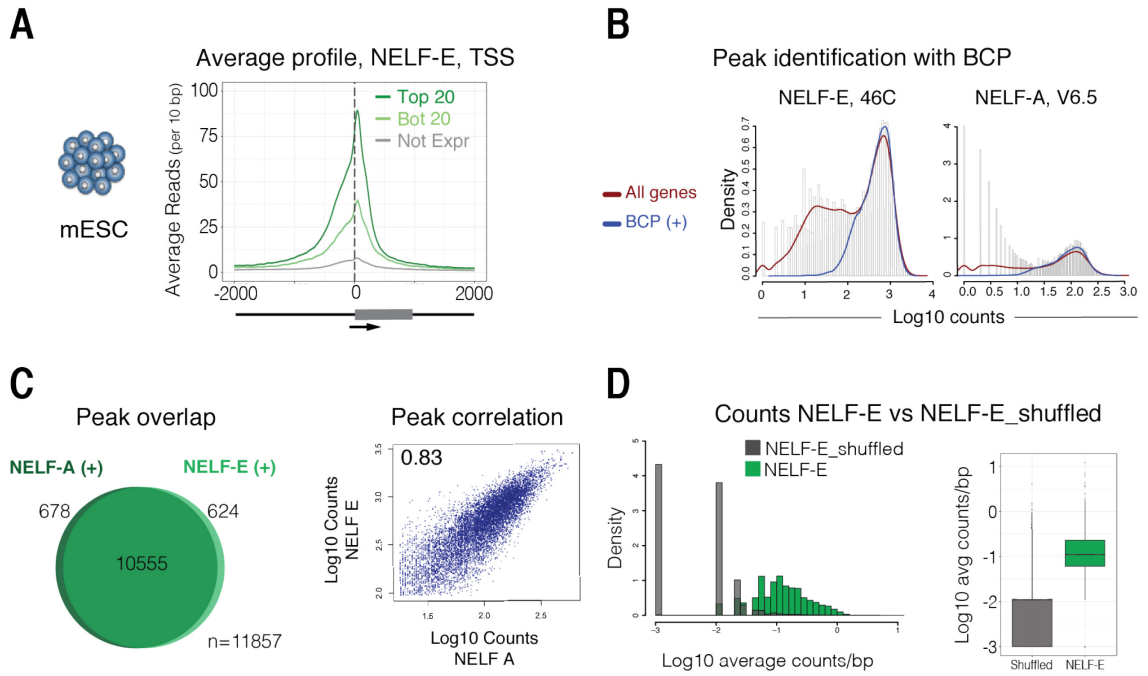

Figure 4.10 Quality control of NELF-E ChIP-seq dataset and comparison with NELF-A published dataset.
A) Average distribution of NELF-E in 4 kb windows centered around the TSS of top 20% (Top 20), bottom 20% (Bot 20) and genes that are not expressed (TPM<1, Not Expr). **B)** Density of reads of NELF-E and NELF-A datasets in promoter windows of all genes compared with genes classified as positive by BCP. **C)** Overlap of BCP positive windows at TSS between NELF-E and NELF-A and Spearman's correlation between overlapping peaks. **D)** Comparison of read density between BCP-positive NELF-E peaks at TSS and shuffled peaks. Boxplot shows higher read count in NELF-E BCP-positive peaks when compared to shuffled peaks.

To further investigate the regulation of RNAPII recruitment and modification at genes that produce circRNAs, I next determined the occupancy and abundance of general transcription factors, RNAPII transcription initiation regulators (CDK7 and 8) and promoter-proximal pausing modulators (NELF-E and CDK9). First, I inspected the occupancy on chromatin of the above-mentioned factors at specific circ- and circ+ genes on the UCSC genome browser, including the circ- gene *Gtf3c3* and the circ+ gene *Rabep1* (**Fig. 4.11**; same genes as in **Fig. 4.3**). General transcription factor TAF1 and transcription initiation regulator CDK7, which phosphorylates CTD residues on S5 and S7, show comparable occupancy at *Gtf3c3* and *Rabep1* genes. However, promoter-proximal pausing modulators

NELF and CDK9 are less enriched at the TSS of circ+ gene *Rabep1* gene than to circ- gene *Gtf3c3*.

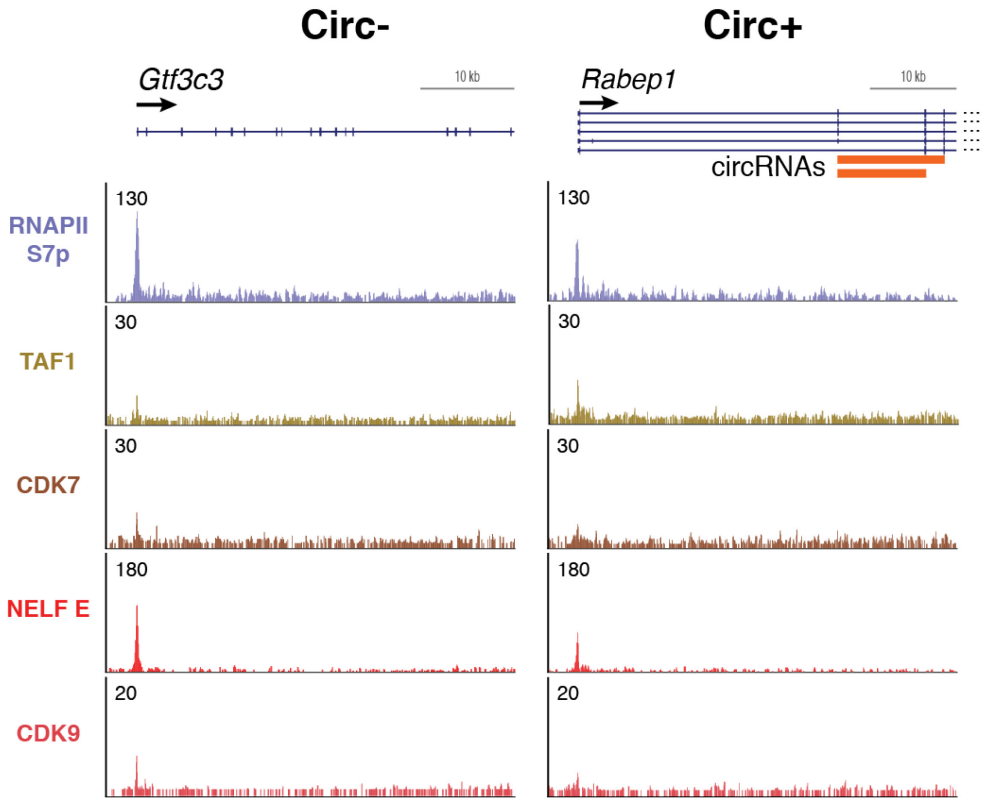

Figure 4.11 Single gene profiles to illustrate the occupancy of factors that modulate transcription at different stages of the transcription cycle in circ- and circ+ genes.
Examples to highlight the differences in enrichment of general transcription factor TAF1, transcription initiation factor CDK7 and promoter-proximal pausing modulators NELF-E and CDK9 between circ- gene *Gtf3c3* and circ+ gene *Rabep1* in mESCs.

These observations prompted an investigation of the occupancy of transcription regulators genome-wide at circ+ and circ-genes in mESC, using the groups of genes defined on section 4.5.1. I first determined the occupancy of general transcription factors TBP and TAF1, which are part of TFIID, a component of the core transcription initiation complex. Circ+ and circ- genes show comparable average profiles and enrichment level of TBP and TAF1 at the TSS, with a slight decrease at circ+ genes when zooming into the E1-I1 border (**Fig. 4.12A**). These results point to RNAPII recruitment being mostly unaffected at circ+ genes. I then examined CDK7 and CDK8, which phosphorylate S5 and S7 and modulate transcription initiation. CDK7 is slightly decreased at the TSS with a stronger decrease at E1-I1 border, while CDK8 is mostly unchanged at the TSS and also

decreased at the E1-I1 border of circ+ genes (**Fig. 4.12B**). The slight depletion of CDK7 and CDK8 at circ+ genes does not seem sufficient to explain the decreased occupancy S5p or S7p at promoter regions of circ+ genes, suggesting that additional factors may contribute to the marked depletion of S5p and S7p at circ+ genes. These results suggest that circ+ and circ- genes have similar overall recruitment of RNAPII to the promoter and transcription initiation.

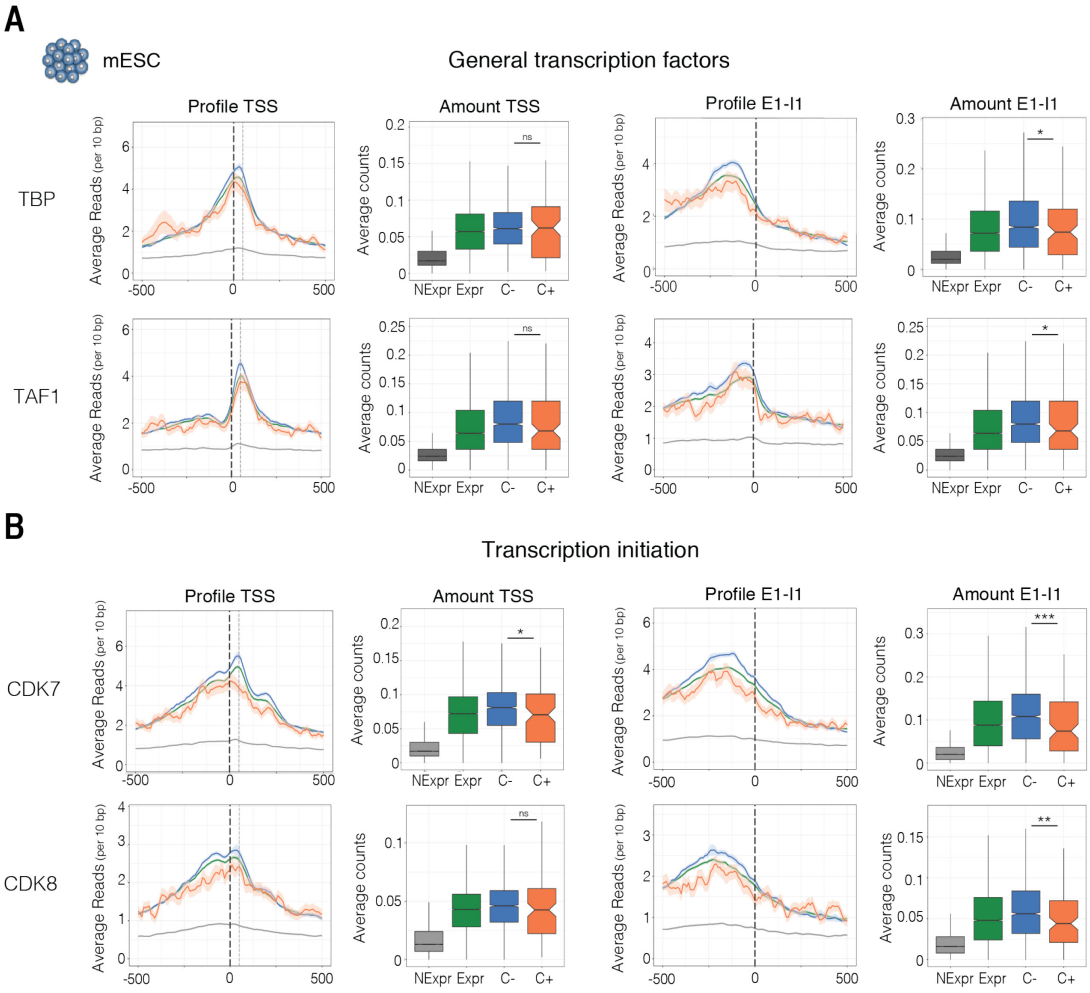

Figure 4.12 Enrichment of general transcription factors and transcription initiation factors at the TSS and E1-I1 border in mESCs.
Average distribution of TBP, TAF1, CDK7 and CDK8 in 1kb windows centered at TSS and E1-I1 border. Boxplots show enrichment at these locations. Significance determined with Wilcoxon rank sum test; * p-value < 0.05, ** p-value < 0.01; *** p-value < 0.001; ns – not significant.

Given that RNAPII recruitment and transcription initiation were generally unchanged at circ+ genes, I asked whether circ- and circ+ genes have different extents of promoter-proximal pausing. To this end, I analyzed the occupancy on chromatin of NELF-A and -E subunits, parts of the NELF complex which, together

with DSIF, pause RNAPII close to the promoter. I also analyzed CDK9, component of the PTEF-b complex, which phosphorylates RNAPII on S2 residues, NELF and DSIF, triggering RNAPII release from the promoter into productive elongation. Remarkably, both NELF subunits and CDK9 are markedly depleted at the TSS of circ+ genes when compared to circ-, an effect that is most striking at the E1-I1 border (**Fig. 4.13**). These results suggest that promoter-proximal pausing is affected at circ+ genes.

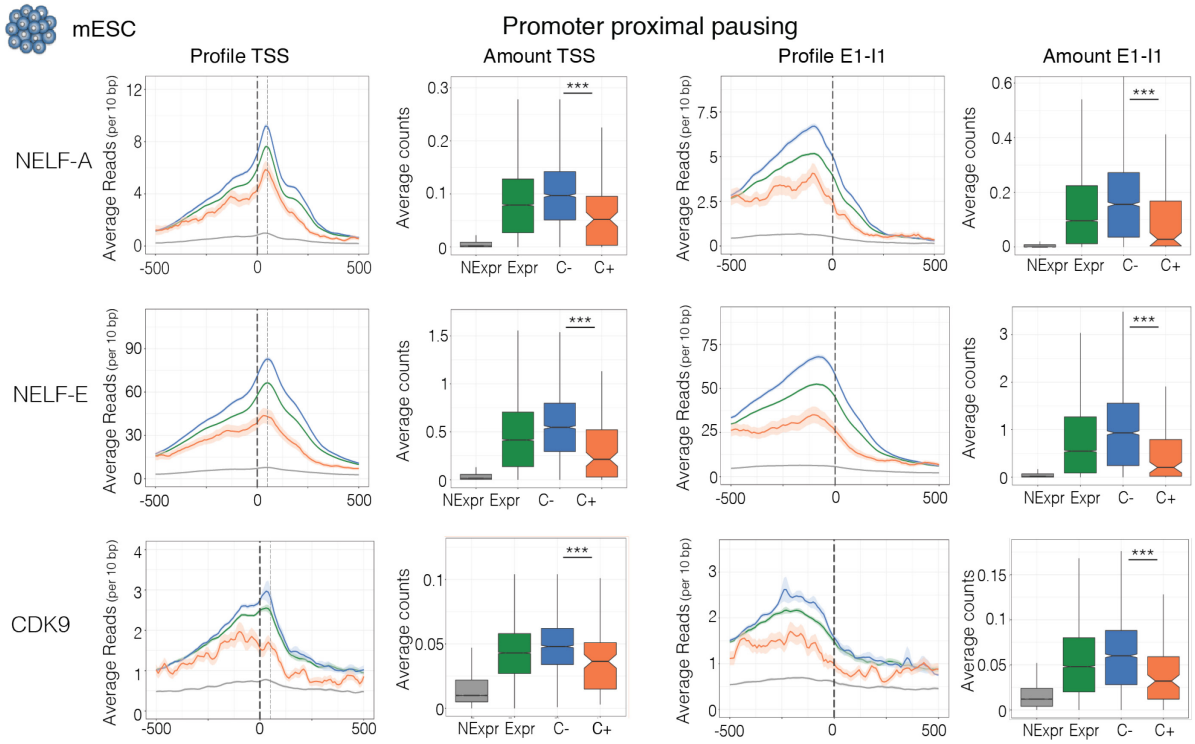

Figure 4.13 Enrichment of promoter-proximal pausing factors at the TSS and E1-I1 border in mESCs.
Average distribution of NELF-A, NELF-E and CDK9 in 1kb windows centered at TSS and E1-I1 border. Boxplots show enrichment at these locations. Significance determined with Wilcoxon rank sum test; *** p-value < 0.001.

In summary, the results present in this chapter show that circ+ genes identified in mESCs have altered co-transcriptional splicing dynamics in mESCs, whereas RNAPII recruitment and transcription initiation do not seem to be greatly affected. Taking into account that circ+ genes have high expression but lower occupancy of S5p and S7p modifications and promoter-proximal pausing factors, it seems that RNAPII transitions from initiation to elongation too quickly without reaching the appropriate levels of S5p and S7p modification, which then likely

contributes to decreased spliceosome recruitment when RNAPII crosses the E1-I1 junction, which then leads to intron 1 temporary retention and opens the opportunity for increased of circRNA formation. To expand from these observations in mESCs, and understand if similar promoter-based mechanisms may underlie circRNA formation in neurons, where circ+ genes are most common, I produced new datasets and investigated published datasets from *in vitro* grown neurons.

4.7 Promoter-based regulation of circRNA biogenesis in neurons

To expand from the analyses in mESCs presented above, I next investigated the occupancy of RNAPII, NELF, and U1 snRNP in dopaminergic neurons differentiated *in vitro* at days 16 and 30. I took advantage of published ChIP-seq for RNAPII S5p and S7p in dopaminergic neurons days 16 and 30 (Ferrai et al. 2017). I also produced new datasets for RNAPII S5p in spinal motor neurons days 2 and 8.

4.7.1 RNAPII S5p and S7p are slightly depleted at circRNA-producing genes in dopaminergic neurons

I started by investigating whether the features of genes that produce circRNAs in dopaminergic neurons. Similarly to mESCs, circRNAs produced at days 16 and 30 most often start at exon 2. To identify a control group of expressed genes that do not make circRNAs but are expressed at similar levels, I selected circ+ and circ- genes for each time-point based on the parameters described in section 4.4. First, I confirmed that the groups of circ- genes selected for days 16 or 30 have comparable expression to circ+ genes in the same time point (**Fig. 4.14**). In contrast, and with the same tendency as found in mESCs (**Fig. 4.5**), circ+ genes have longer intron 1 than circ- genes (18.66 kb compared with 2.45 kb at day 16 and 19.86 kb compared with 2.57 kb at day 30), and slightly longer exon 1 (233 bp vs 166 bp at day 16 and 241 bp vs 163 bp at day 30) and exon 2 (135 bp vs

119 bp at day 16 and 129 bp vs 121 bp at day 30) (**Fig. 4.14).** Expressed genes are shown for comparison.

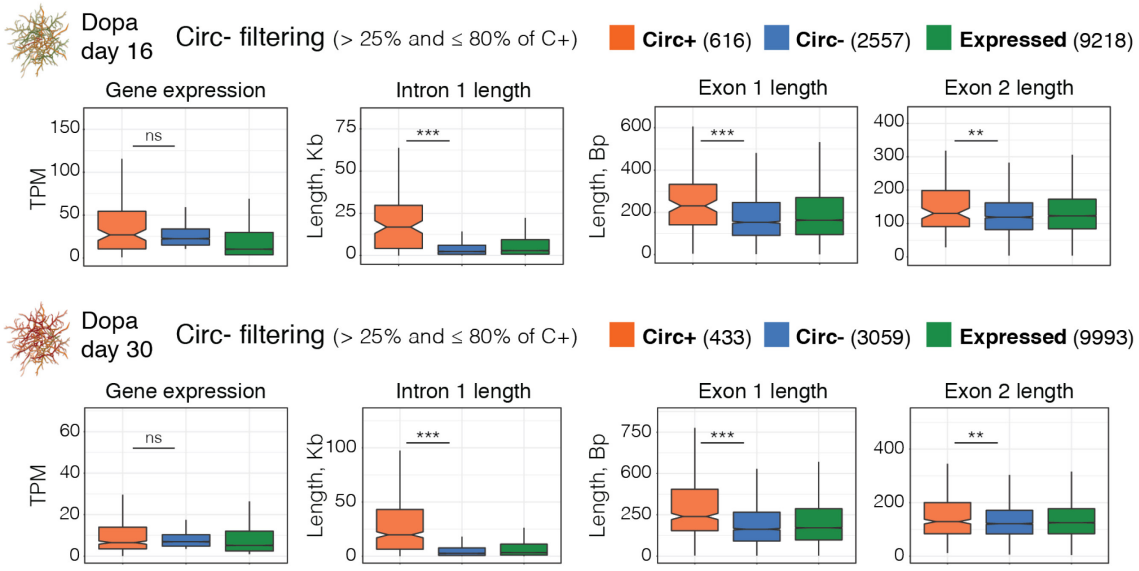

Figure 4.14 Circ- gene filtering to match circ+ expression in dopaminergic neurons days 16 and 30.
Boxplots showing gene expression levels, intron 1 and exons 1 and 2 length for circ+, circ- and expressed genes in mESCs. Circ- genes were selected with a threshold from 25% to 80% of circ+ expression. Significance determined with Wilcoxon rank sum test; ** p-value < 0.01; *** p-value < 0.001; ns – not significant.

I next took advantage of published ChIP-seq datasets for RNAPII S5p and S7p for dopaminergic neurons days 16 and 30 (Ferrai et al. 2017) and determined their occupancy at the TSS and E1-I1 border of circ+ and circ- genes. Expressed and not expressed genes were included as controls. As for mESC, in day 16 and day 30 dopaminergic neurons, both S5p and S7p are found slightly decreased at the promoter of circ+ genes compared to circ-, an effect that is more pronounced at the E1-I1 border (**Fig. 4.15A, B**). A negative control with the unspecific antibody Dig in day 16 dopaminergic neurons does not show enrichment at the TSS, as expected (**Fig. 4.15C**). Although the depletion of S5p and S7p in dopaminergic neurons is not as marked as in mESCs, the levels of S5p may be sufficiently decreased to result in altered spliceosome recruitment, and the decreased levels of S7p may still reflect decreased promoter-proximal pausing.

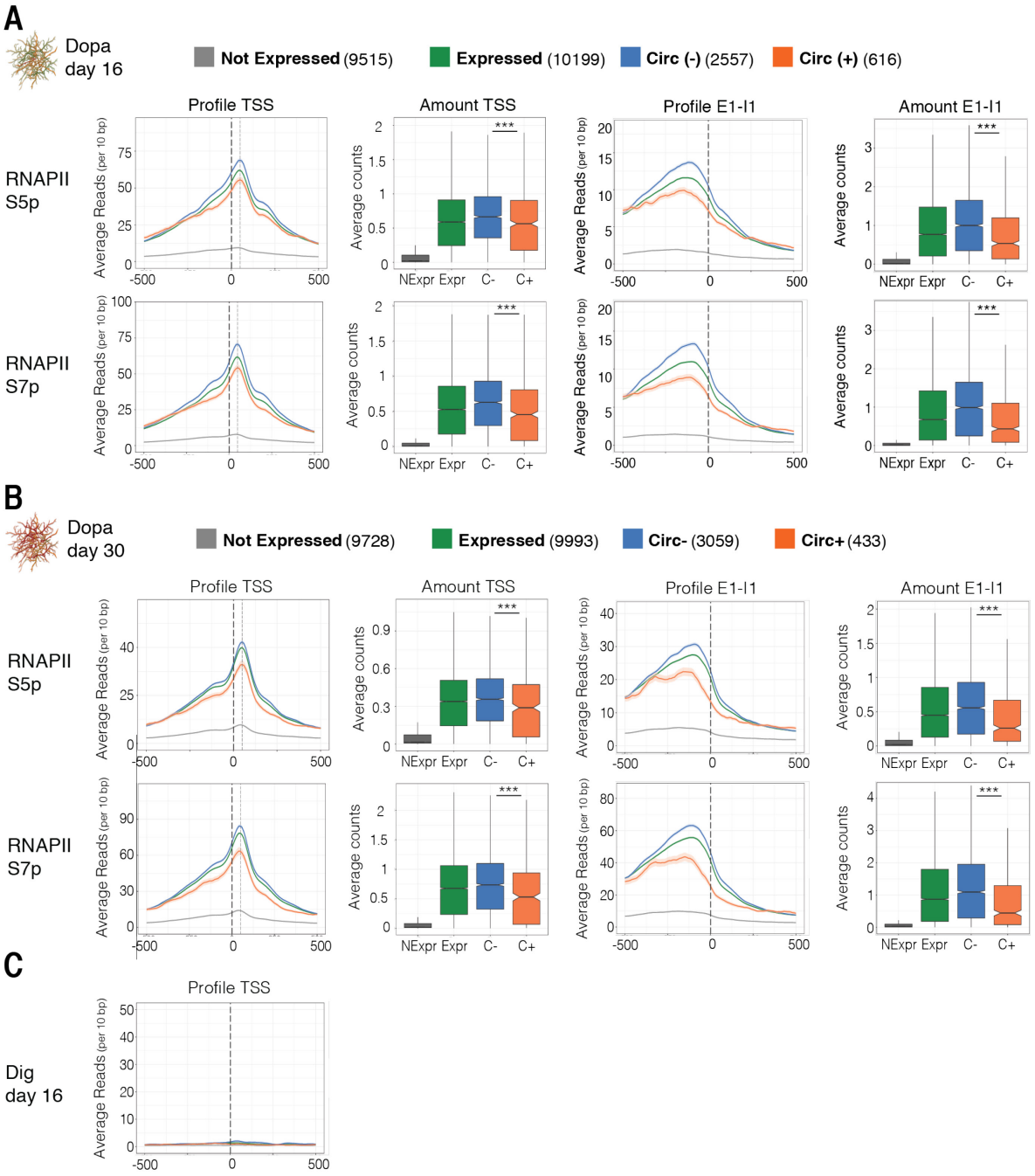

Figure 4.15 Enrichment of RNAPII S5p, S7p and Dig at the TSS and E1-I1 border in dopaminergic neurons days 16 and 30.
Average distribution of RNAPII S5p, S7p Dig in 1kb windows at TSS and E1-I1 border of days **A)** 16 and **B)** 30 dopaminergic neurons. Boxplots show enrichment at these locations. **C)** Negative control Dig does not show enrichment at TSS. Significance determined with Wilcoxon rank sum test; *** p-value < 0.001.

4.7.2 NELF is depleted at circRNA-producing genes in dopaminergic neurons

To further explore the possibility of altered RNAPII release from the promoter at circ+ genes in dopaminergic neurons, I next mapped NELF-E occupancy on

chromatin using day 16 dopaminergic neurons. I started by confirming the quality of the NELF-E ChIP-seq dataset. NELF-E is enriched at the promoter of expressed genes in dopaminergic neurons and its enrichment level correlates with gene expression (top 20% vs bottom 20% vs Not expr, **Fig. 4.16A**). After BCP analyses to determine NELF-E peaks, I determined enrichment at gene promoters (**Fig. 4.16B**), and I confirmed that the NELF-E positive windows have higher read counts when compared to shuffled BCP windows, indicating that the classification of NELF-E positive windows is specific (**Fig. 4.16C**).

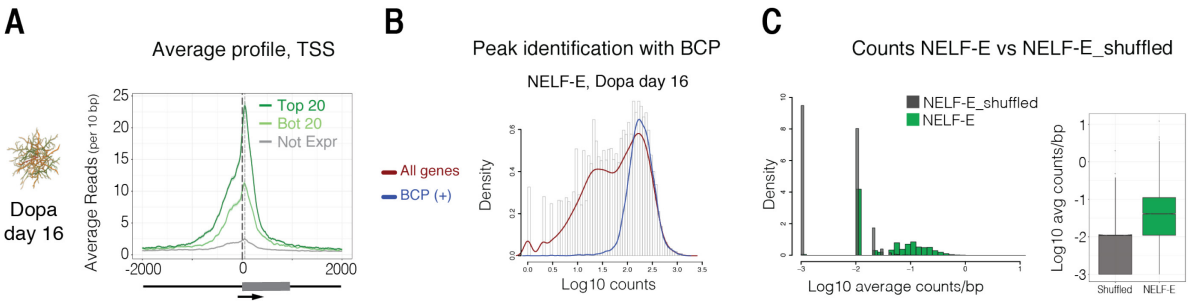

Figure 4.16 Quality control of NELF-E ChIP-seq in dopaminergic neurons day 16.
A) Average distribution of NELF-E in 4 kb windows around centered at the TSS of top 20%, bottom 20% and genes that are not expressed (TPM<1). **B)** Density of reads of NELF-E in promoter windows of all genes compared with genes classified as positive by BCP. **C)** Comparison of read density between BCP-positive NELF-E peaks at TSS and shuffled peaks. Boxplot shows higher read count in NELF-E BCP-positive peaks when compared to shuffled peaks.

I next determined the occupancy of NELF-E on chromatin of circ+ and circ- genes at dopaminergic neurons day 16. In agreement with the observations made in mESCs, I found that NELF is also markedly depleted at the TSS and E1-I1 border of circ+ genes when compared to circ- (**Fig. 4.17**). These results indicate that RNAPII release from the promoter is altered at circ+ genes in dopaminergic neurons day 16 as in mESCs, suggesting that lower NELF-E recruitment is likely to be a general feature of circ+ genes. Altogether, slightly decreased occupancy of RNAPII S5p and S7p with marked depletion of NELF complex points towards increased RNAPII release from promoter-proximal pausing contributing to circRNA formation also in dopaminergic neurons.

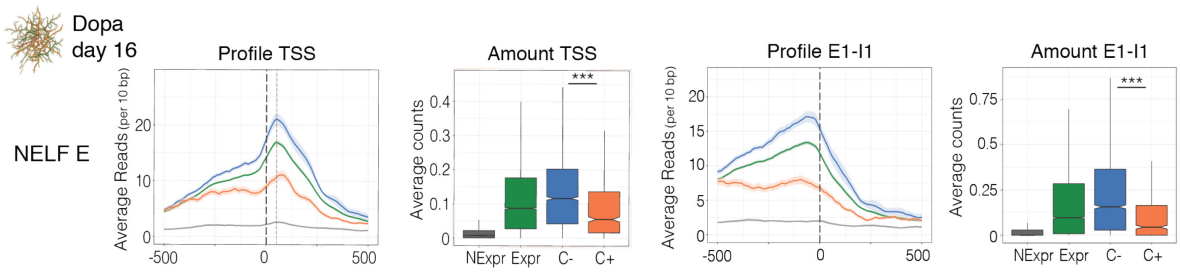

Figure 4.17 Enrichment of promoter-proximal pausing factor NELF-E at the TSS and E1-I1 border in day 16 dopaminergic neurons.
Average distribution NELF-E in 1kb windows centered at TSS and E1-I1 border. Boxplots show enrichment at these locations. Significance determined with Wilcoxon rank sum test. *** p-value < 0.001.

4.7.3 RNAPII S5p is depleted at circRNA-producing genes in spinal motor neurons

To further explore the generality of my observations in mESCs and dopaminergic neurons, I next investigated whether increased if RNAPII release from the promoter contributes to circRNA production in a distinct neuronal subtype. To address this, I mapped and the occupancy of RNAPII S5p in spinal motor neurons at days 2 and 8 and applied the previous strategy. As expected, RNAPII S5p is enriched at the TSS of expressed genes and its enrichment correlates with expression levels (top 20% vs bottom 20%) and S5p is not enriched at genes that are not expressed (**Fig. 4.18A**). Further analyses with BCP confirm that S5p positive windows are present at the TSS of many genes (**Fig. 4.18B**). Finally, S5p positive windows show higher read counts when compared to shuffled positive windows, indicating that S5p enrichment is specific and that these datasets are good for further use (**Fig. 4.18B).** I then selected circ+, circ-, expressed and not expressed genes for days 2 and 8 to further determine enrichment of S5p at these genes. As is shown in **Fig. 4.19**, circ+ and circ- genes have comparable gene expression levels, and as found previously for mESC and dopaminergic neurons, circ+ have longer intron 1 and exons 1 and 2.

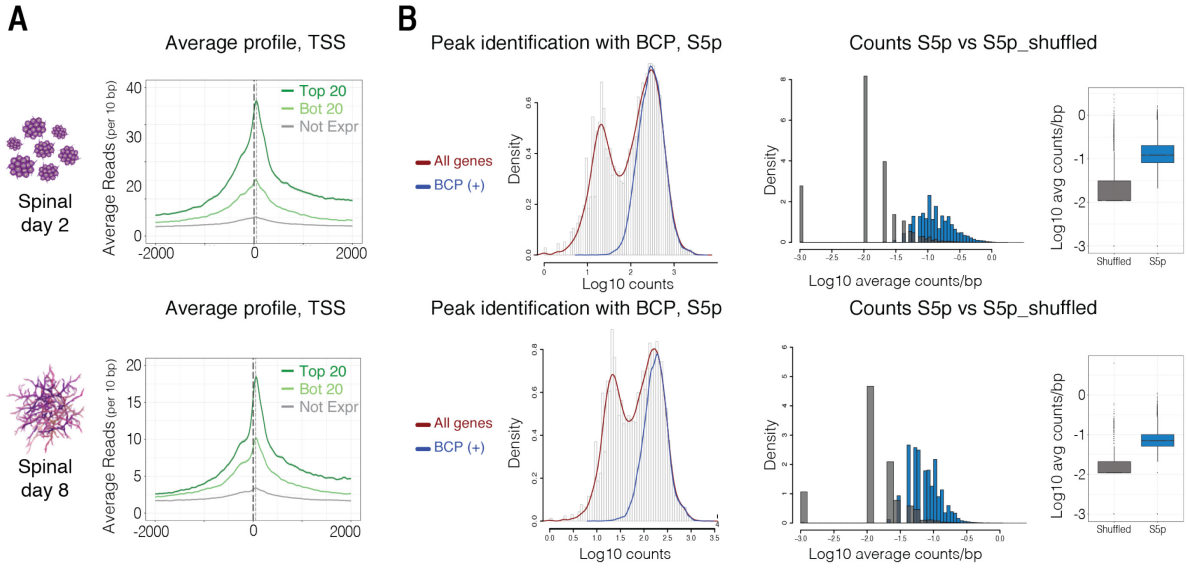

Figure 4.18 Quality control of RNAPII S5p ChIP-seq in spinal motor neurons days 2 and 8.
A) Average distribution of S5p in 4 kb windows centered around the TSS of top 20%, bottom 20% and genes that are not expressed (TPM<1). B) Left panel: density of reads of S5p dataset in promoter windows of all genes compared with genes classified as positive by BCP. Right panel: comparison of read density and counts between BCP-positive S5p peaks at TSS and shuffled peaks. Boxplot shows higher read count in S5p BCP-positive peaks when compared to shuffled peaks.

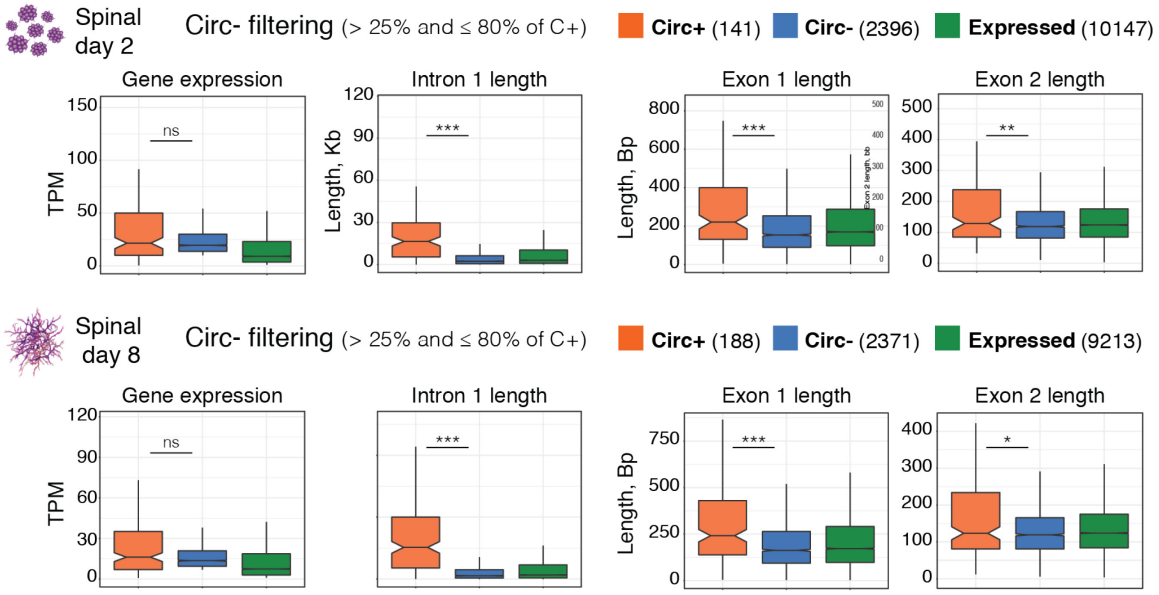

Figure 4.19 Circ- gene filtering to match circ+ expression in spinal motor neurons days 2 and 8.
Boxplots showing gene expression levels, intron 1 and exons 1 and 2 length for circ+, circ- and expressed genes in mESCs. Circ- genes were selected with a threshold from 25% to 80% of circ+ expression. Wilcoxon rank sum test. *** p-value < 0.001; ns – not significant.

I next determined RNAPII S5p occupancy at the TSS and E1-I1 border of all gene groups selected. Results show that S5p is slightly depleted at the promoter of circ+ genes compared to circ- of day 2 spinal motor neurons, an effect that is

more profound at the E1-I1 border (**Fig. 4.20A**) Interestingly, day 8 spinal motor neurons show a marked depletion of S5p at the TSS and E1-I1 border (**Fig. 4.20B**).

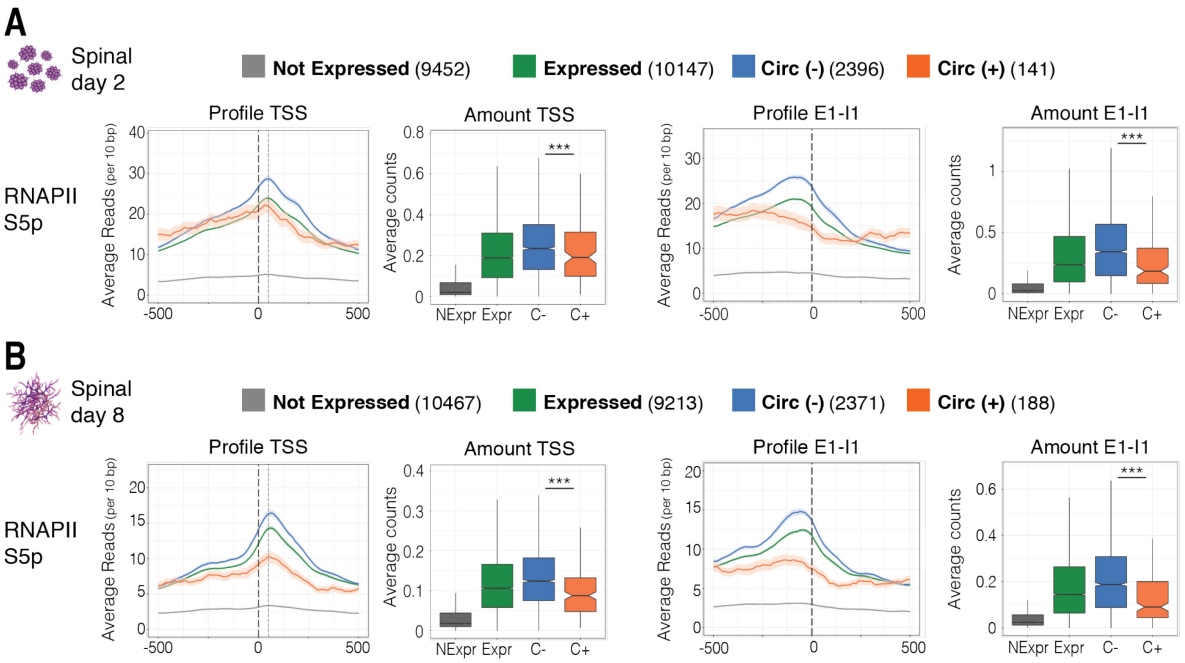

Figure 4.20 Enrichment of RNAPII S5p at the TSS and E1-I1 border in spinal motor neurons days 2 and 8.
Average distribution S5p in 1kb windows at TSS and E1-I1 border in **A)** days 2 and **B)** 8 spinal motor neurons. Boxplots show enrichment at these locations. Wilcoxon rank sum test. *** p-value < 0.001

Altogether, the results show that, similarly to mESCs, RNAPII S5p and S7p enrichment at circ+ genes is decreased in dopaminergic and spinal motor neurons. Taken together with the NELF complex being markedly decreased in dopaminergic neurons and mESCs, this supports the notion that altered RNAPII release from promoter-proximal pausing at circ+ genes in mESCs and neurons is a mechanism that may lead to circRNA formation.

4.8 Exploring the mechanism of RNA polymerase II promoter escape in the regulation of circRNA formation

4.8.1 Knockdown of NELF complex in mESC

According to the results presented above, impaired promoter-proximal pausing may trigger the release of RNAPII from the promoter into productive elongation and increase circRNA formation. To test if RNAPII release from the promoter plays a role in circRNA formation, I set out to perturb promoter-proximal pausing mechanisms by knocking down NELF, DSIF, or both complexes. To interfere with promoter release, I designed siRNAs for depletion of NELF and SPT5 using transient transfections in mESCs. Amongst the siRNAs tested, the knockdown of NELF-A was successful (**Fig. 4.21A, B**). I first studied whether NELF-A knockdown interfered with ESC growth, morphology of expression of pluripotency and differentiation markers. *In vivo* microscopy images of mESCs transfected with NELF-A siRNAs (NELF-A KD) or scrambled (Scrambled) siRNAs show no detectable changes in cell morphology (**Fig. 4.21C**), suggesting that mESCs retain their state of pluripotency.

To investigate gene expression, I collected RNA from two biological replicates and produced total RNA-seq libraries. After sequencing, I quantified gene expression in TPMs, from scrambled and NELF-A KD, and found that replicates correlate very well (Pearson's correlation of 0.95 and 0.96, respectively) (**Fig. 4.21D**). To further understand the effect of NELF-A knockdown, I investigated the transcript expression levels of pluripotency transcription factors Oct4 and Nanog, early differentiation markers Hes5 and Fgf5, and subunits of the NELF, DSIF, the largest subunit of RNAPII (Rpb1), U1 snRNP and U2 snRNP complexes (**Fig. 4.21E**). While *NELF-A* expression is markedly reduced in both knockdown replicates compared to scrambled siRNA transfection (~70% in both replicates), the expression of other subunits constituting the NELF complex is unchanged. Similarly, the expression of DSIF subunits *SPT4/5*, and of transcription and splicing components show no major differences. The only exception is the U2 snRNP small subunit, *U2-A*, whose expression is ~30% lower in both knockdown biological replicates when compared to scrambled. Furthermore, expression of pluripotency markers *Nanog* and *Pou5f1* and early differentiation markers *Fgf5* and *Hes5* is unaffected, suggesting that mESCs maintain pluripotency upon NELF-A knockdown.

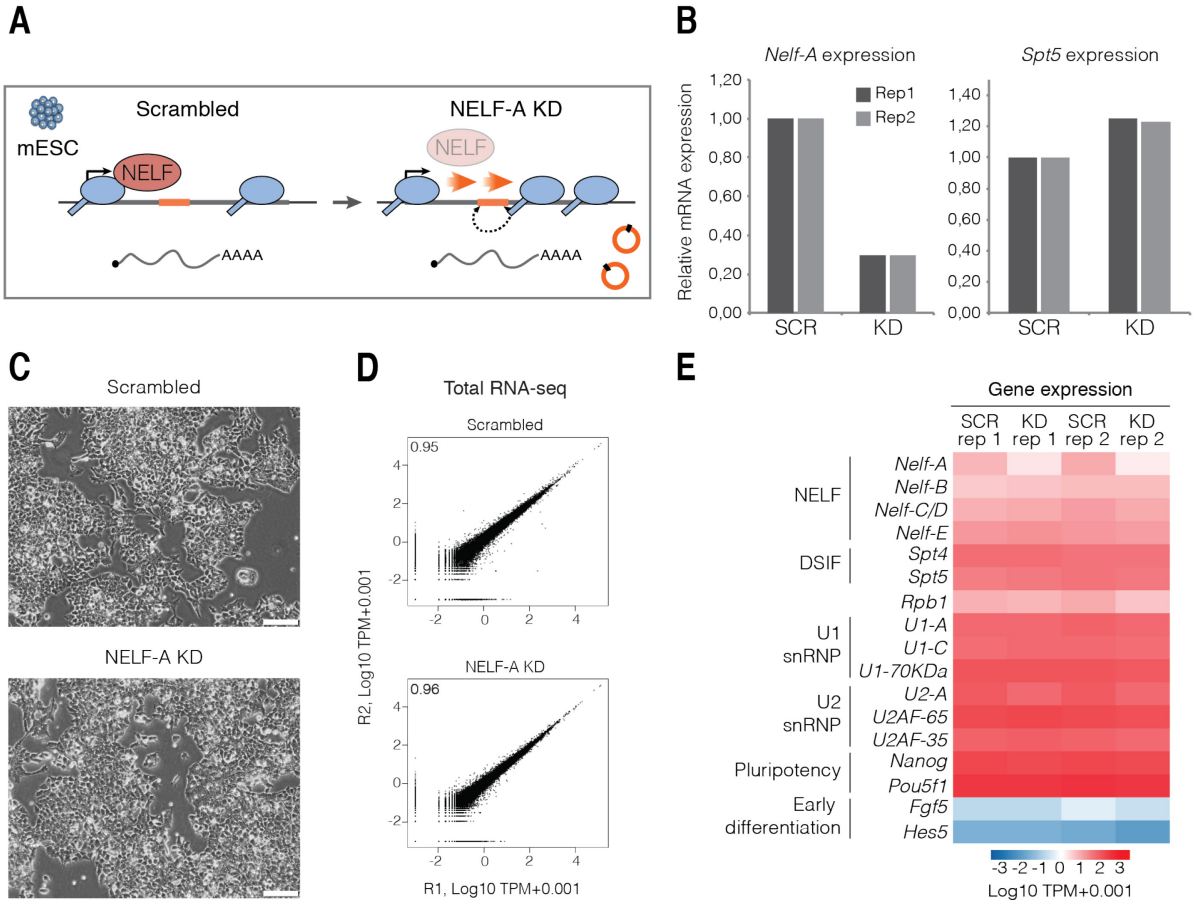

Figure 4.21 Experimental design and characterization of NELF knockdown in mESCs.
A) Schematic illustration of NELF knockdown experimental design. mESCs were transfected with a scrambled siRNA without cellular targets or with siRNA targeting NELF-A subunit. **B)** RNA expression of *Nelf-A* and *Spt5* genes in scrambled siRNA (SCR) or NELF-A knockdown (KD) quantified by qPCR in biological replicates 1 and 2. Relative levels are first normalised to *Actb* gene and then to expression in SCR. **C)** *In vivo* light microscopy of mESC cultures transfected with scrambled or NELF-A siRNAs. Bar corresponds to 100 μm. **D)** Pearson's correlation of gene expression for total RNA-seq libraries from two biological replicates of scrambled and NELF-A knockdown (KD). **E)** Expression of several relevant genes in both biological replicates of scrambled (SCR) and NELF-A knockdown (KD).

To test whether NELF-A knockdown has a general effect on gene expression, I calculated the amount of linear transcripts from the total RNA-seq dataset between scrambled and knockdown samples, and compared the distribution and cumulative density of expressed genes (TPM ≥ 1). The two biological replicates were analysed in parallel. Bulk analysis and cumulative density plots show no detectable difference between the distribution of gene expression values in scrambled and knockdown datasets, and that expressed genes in both samples overlap extensively (10839 and 11280 genes for biological replicates 1 and 2,

respectively) (**Fig. 4.22**). Thus, NELF-A knockdown does not greatly impact gene expression.

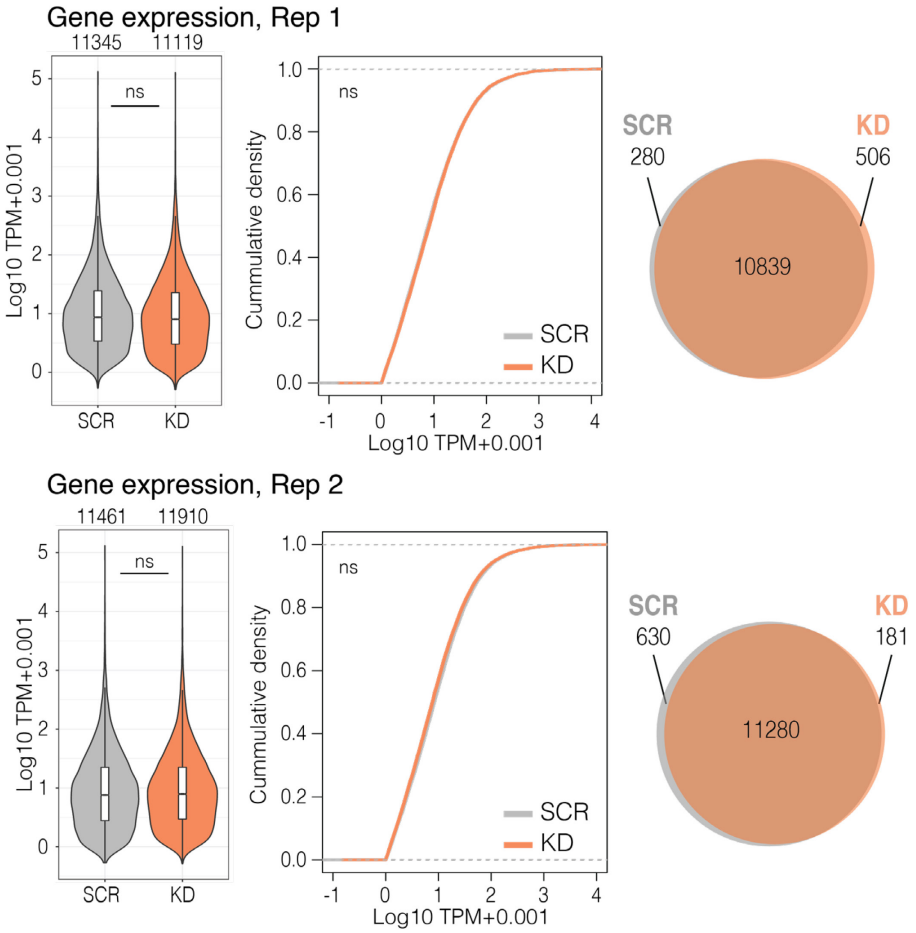

Figure 4.22 Effect of NELF-A knockdown on gene expression.
Violin plots, cumulative density plots and overlaps between expressed genes (TPM ≥ 1) in SCR or KD samples. The number of expressed genes in SCR or KD samples is indicated above the violin distributions. Individual biological replicates were analysed in parallel. Significance of bulk gene expression was determined with a two-tailed t-test and for cumulative density with the Kolmogorov–Smirnov test. ns-not significant.

4.8.2 NELF depletion increases the number of circRNAs and genes producing circRNAs

After evaluating the effects of NELF-A knockdown on expression of mRNAs, I set out to explore its effects on circRNA production. CircRNAs were quantified in both biological replicates of scrambled and knockdown samples (Petar Glažar, laboratory of Nikolaus Rajewsky), and a total of 1794 circRNAs were detected.

The left panel of **Fig. 4.23** shows the number of circRNAs detected per replicate of mESC transfected with scrambled or NELF-A siRNAs and the previous results from untransfected mESC for comparison. Detection of circRNAs between biological replicates is quite variable since most circRNAs are detected uniquely in one of the two replicates (Unique R1 or R2 and **Fig. 3.5**).

We detect a total of 1262 and 874 circRNAs in the NELF-A and scrambled siRNA-treated samples, with 209 and 154 circRNAs being found in both biological replicates (Robust). I next examined the number of genes which produce the circRNAs identified. Accordingly, most genes produce circRNAs detected in individual biological replicates (Unique R1 or R2) and fewer genes produce circRNAs that are robustly detected (Robust) (**Fig. 4.23A**, right panel). I also identified some genes which produce circRNAs that are uniquely detected in each replicate (Non-robust R1+R2). In total, 667 genes produce circRNAs in scrambled samples, where 139 genes produce robustly detected circRNAs, numbers which are comparable to mESCs (667 vs 690 and 139 vs 115 genes, respectively). When analyzing the knockdown samples, 927 genes produce all circRNAs detected and 178 genes produce robustly detected circRNAs.

To determine whether the difference between the number of genes that produce circRNAs in scrambled and knockdown samples is significant, I compared the number of genes producing circRNAs with the number of expressed genes that do not produce circRNAs in scrambled or knockdown samples and applied the Fisher's exact test with a confidence interval of 0.95. I performed this analysis for all genes producing circRNAs and separately for genes that produce robustly detected (Tables 4.1 and 4.2). These results show that NELF-A knockdown significantly increases the number of genes producing circRNAs compared with the control experiment using scrambled siRNA, both for all circRNA-producing genes and for genes producing robust circRNAs detected in two replicates. The same analysis was conducted for individual biological replicates with a similar outcome (not shown). These results support the hypothesis that NELF-A knockdown increases the likelihood of genes producing circRNAs.

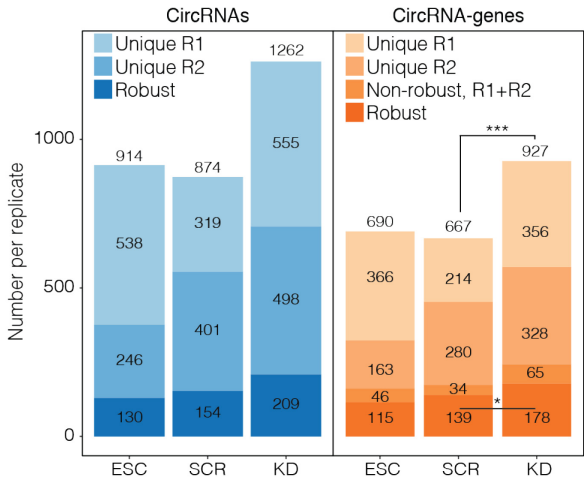

Figure 4.23 Effect of NELF-A knockdown on the number of circRNAs and circRNA-producing genes.
Bar plots showing number of circRNAs (blue) and circRNA-producing genes (orange) found in individual biological replicates (Unique R1, Unique R2) or in both (Robust) in mESCs, SCR and NELF-A KD. Statistical significance between SCR and KD was calculated with Fisher's exact test as is shown on Tables 4.1 and 4.2. * p-value < 0.05, *** p-value < 0.001.

Table 4.1 Comparison of all genes producing circRNAs in scrambled or NELF-A knockdown samples.
Values represent number of genes producing circRNAs (circ+) or genes not producing circRNAs that are expressed (circ-) in scrambled or knockdown samples. Significance was determined with Fisher's exact test.

	Circ+ genes	Circ- genes	Total
NELF-A KD	927	10543	11470
Scrambled	667	10925	11592
Total	1594	21468	23062

Table 4.2 Comparison of genes producing robustly detected circRNAs in scrambled or NELF-A knockdown samples.
Values represent number of genes producing circRNAs (circ+) or genes not producing circRNAs that are expressed (circ-) in scrambled or knockdown samples. Significance was determined with Fisher's exact test.

	Circ+ genes	Circ- genes	Total
NELF-A KD	178	11246	11424
Scrambled	139	11438	11577
Total	317	22684	23001

4.8.3 NELF depletion does not affect the amount of circRNAs produced

I next asked whether the abundance of circRNAs is affected after NELF-A knockdown. To this end, I first determined which circRNAs and circRNA-producing genes are common in both SCR and KD or specifically detected in SCR or KD (**Fig. 4.24**). For all circRNAs detected in biological replicate 1 or 2, the highest number of circRNAs or circRNA-producing genes is detected only in KD samples (595 and 524 circRNAs and 394 and 360 genes, respectively). Fewer circRNAs or circRNA-producing genes are detected only in SCR sample for biological replicates 1 or 2 (304 and 372 circRNAs and 182 and 242 genes, respectively). Finally, some circRNAs or circRNA producing genes are common in both SCR and KD (169 and 183 circRNAs and 205 and 211 genes, respectively). A similar pattern is observed for robustly detected circRNAs or genes which produce these circRNAs: 121 circRNAs and 91 genes are detected only in KD, 66 circRNAs and 52 genes are detected only in SCR and 88 circRNAs or 87 genes are common in both samples.

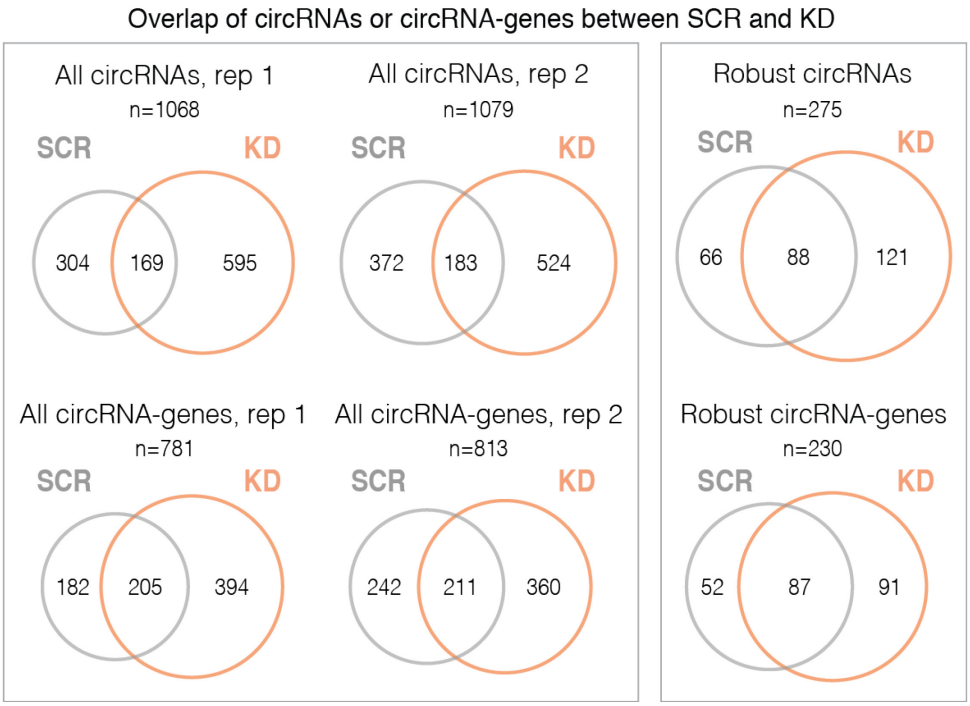

Figure 4.24 Overlaps in scrambled and knockdown samples.

Overlaps of circRNAs or circRNA-producing genes found in scrambled (SCR) and knockdown (KD) individual biological replicates or common in both replicates.

After determining which circRNAs are produced specifically in NELF-A knockdown and scrambled samples, I split all circRNAs into three groups: detected only in scrambled (SCR-only), only in NELF-A knockdown (KD-only) or both (SCR+KD). This terminology is used when referring to these groups of genes in the following sections of the thesis. To compare the expression level of circRNAs, I used the circ-to-linear ratio metric, in each biological replicate and for circRNAs robustly detected in both replicates. Cumulative density graphs of circ-to-linear ratio show that SCR-only circRNAs have higher expression in SCR when compared to KD in both biological replicates, as expected (**Fig. 4.25A**). Conversely, KD-only circRNAs show higher expression in KD when compared to SCR. When analyzing SCR+KD circRNAs, circRNAs tend to be less expressed in knockdown in biological replicate 1 when compared to scrambled, whereas in biological replicate 2 there is no difference. This suggests that the effect of NELF-A knockdown on circRNA expression is either low or not consistent between biological replicates. Similar results were obtained when analyzing robustly detected circRNAs. SCR-only circRNAs are more expressed in scrambled, whereas KD-only circRNAs are more expressed in knockdown (**Fig. 4.25B**). Again, SCR+KD circRNAs tend to be less expressed in the knockdown sample for biological replicate 1 but show unchanged expression in replicate 2.

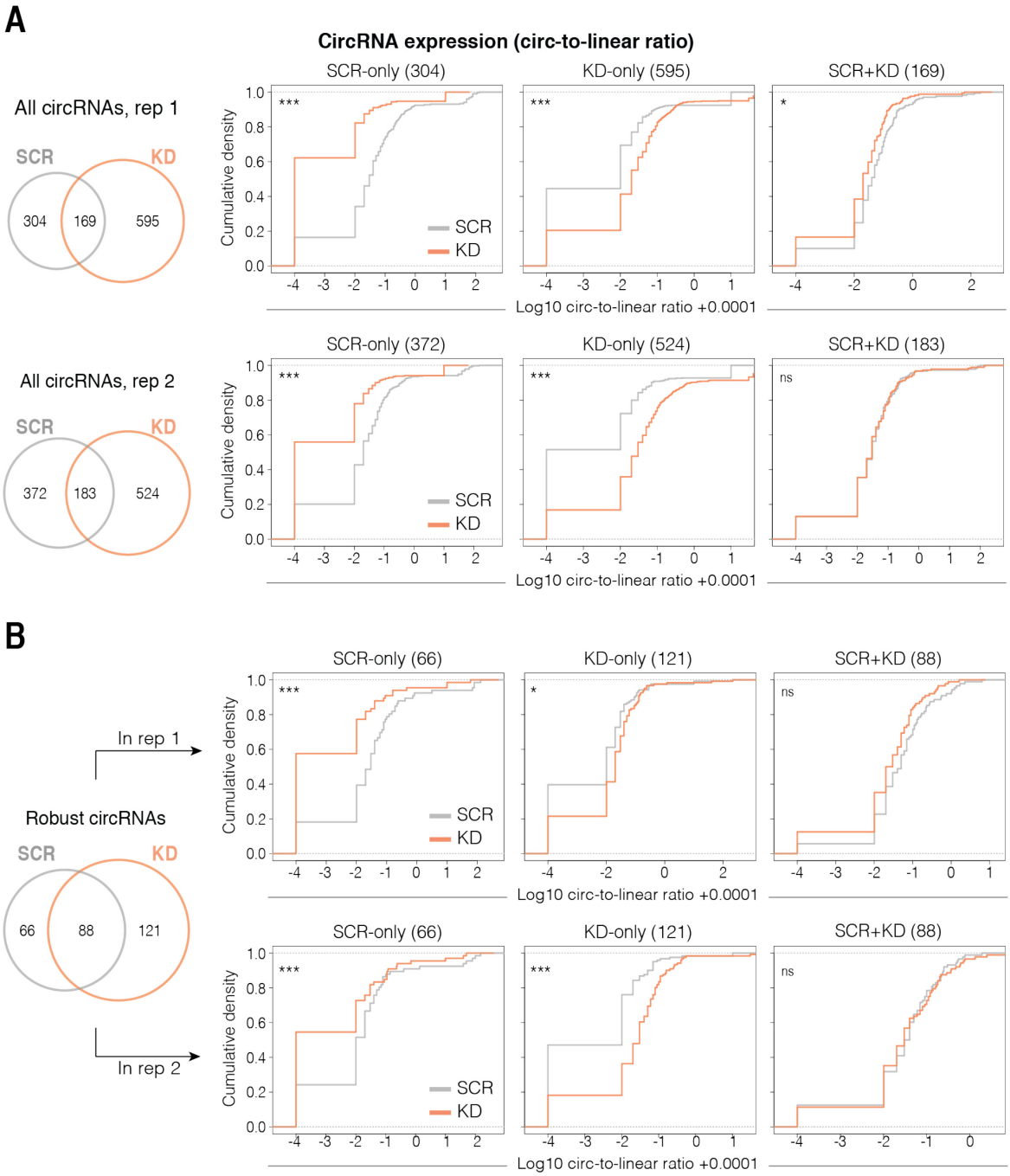

Figure 4.25 Effect of NELF-A knockdown on circRNA expression.
A) Cumulative density plots of circRNA expression in SCR and KD samples for individual biological replicates comparing circRNAs detected only in scrambled (SCR specific), only in KD (KD specific) or common in both (SCR+KD). **B)** Same analysis as in A) with robustly detected circRNAs. Significance determined with the Kolmogorov–Smirnov test. *p-value < 0.05, ** p-value < 0.01; *** p-value < 0.001, ns – not significant.

Altogether, results in this section suggest that NELF-A knockdown leads to an increase in the numbers of circRNAs detected and of genes producing circRNAs, but does not have a detectable impact the expression of circRNAs relative to their linear isoform. Though analysis for all circRNAs detected in individual biological

replicates was performed, for the sake of simplicity the following analyses show results for robust circRNAs only.

4.8.4 Most genes producing circRNAs after NELF knockdown produce circRNAs in mESCs and during neuronal differentiation

After determining that NELF-A knockdown increases the likelihood of genes producing circRNAs, I asked which genes are these and what are their features. I started by comparing KD-only, SCR-only or SCR+KD genes with genes that produce circRNAs throughout dopaminergic and spinal motor neuronal differentiations. I represented circRNA expression per gene and linear gene expression in a colored heatmap where genes were: firstly, ranked by their overlap with genes that produce robust or unique circRNAs during differentiation; secondly, by decreasing expression of KD-only, SCR+KD or SCR-only genes; and finally, by decreasing expression at all time-points of both differentiations. Values corresponding to biological replicate 1 are shown in this figure (**Fig. 4.26**). From the 230 genes that produce robust circRNAs after NELF-A knockdown, 168 produce robust circRNAs at least one time-point of both differentiations ("Robust in differentiation"). The largest group consists SCR+KD genes: many of these genes produce circRNAs in mESCs (49 out of 79, ~ 62%) and are also detected in all other time-points with an expression pattern that is almost ubiquitous. KD-only genes are also detected at all time-points of both differentiations, where ~25% of the genes produce circRNAs in mESCs (15 out of 60) and most produce circRNAs at day 16 dopaminergic neurons (46 out of 60, ~ 77%). Finally, SCR-only genes also produce circRNAs at all time-points, where ~ 38% are common with mESCs (11 out of 29) and ~ 76% are common with day 16 dopaminergic neurons (22 out of 29). When observing the RNA expression from these genes, expression levels appear very similar between KD, SCR and mESCs samples and most genes seem expressed throughout differentiation. Additionally, of the 230 genes producing robust circRNAs in NELF-A KD experiments, 57 genes were also found to produce uniquely detected circRNAs during differentiation (Unique in differentiation). Of these, 28 genes are detected only in KD, 22 only in

SCR and 8 are common in SCR and KD. All groups of genes produce circRNAs during both neuronal differentiations, with highest overlap in dopaminergic neurons days 16 and 30. RNA expression from these genes also does not seem to differ between SCR, KD and mESCs and these genes appear to be mostly expressed throughout differentiation. Finally, only 5 genes that produce circRNAs in NELF-A KD experiments do not produce circRNAs during differentiation ("*de novo*" genes). This shows that the vast majority of genes producing circRNAs in NELF-A KD experiments also produce circRNAs at various time-points during neuronal differentiation. To further understand which genes produce circRNAs only in SCR, KD or both, I performed gene ontology analyses using all genes as background which did not yield enrichment of specific gene groups (not shown).

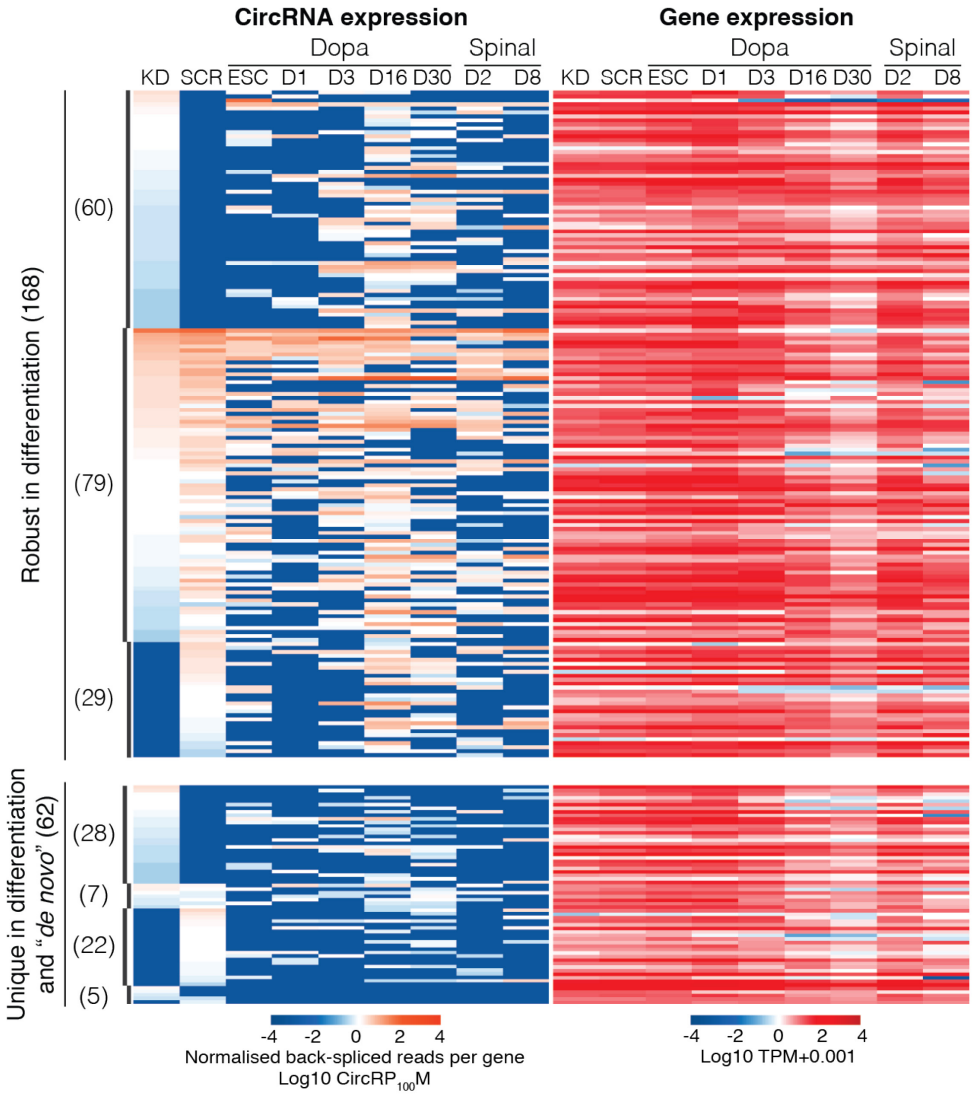

Figure 4.26 Expression patterns of circRNA-producing genes from NELF-A KD experiments in dopaminergic and spinal motor neuron differentiation.
Heatmaps of circRNA expression per gene (CircRP100M) and gene expression (TPM) in SCR, KD, and both neuronal differentiations. Top panel contains genes that produce robustly detected circRNAs in KD and SCR that are common in both differentiations (Robust in differentiation). Bottom panel contains genes that produce robustly detected circRNAs in KD and SCR common with genes that produce circRNAs detected in a single biological replicate of both differentiations (Unique in differentiation) and circRNA-genes that are found only in SCR and KD samples ("De novo"). Values correspond to biological replicate 1.

4.8.5 Genes producing circRNAs upon NELF depletion tend to be highly expressed

Results so far show that NELF-A knockdown leads to an increase in the number of genes producing circRNAs without detectable effects on the level of expression of circRNAs, suggesting that the new circ+ genes detected in the NELF-A knockdown are more sensitive to NELF-A knockdown. The increased detection of

larger number of genes producing circRNAs could suggest that some genes have specific properties that make them more sensitive to NELF levels or alternatively that depletion of NELF-A leads to a general deregulation of promoter-proximal pausing across a larger group of genes.

To explore whether the detection of new circ+ genes could result from specific gene features, I compared the expression and structural features of circ+ genes found only in the NELF-A siRNA knockdown, in cells treated with scrambled siRNAs and in both treatments. I started by comparing the linear gene expression of these three groups of genes with expressed genes (TPM ≥ 1). As expected of genes producing circRNAs, genes that produce circRNAs in both NELF-A and scrambled siRNA treatments (SCR+KD) are more highly expressed than expressed genes in cells treated with NELF-A or scrambled siRNAs (**Fig. 4.27A, B**). Genes that produce circRNAs specifically in NELF-A knockdown are significantly more highly expressed than expressed genes in both the cells treated with NELF-A or scrambled siRNAs. SCR-only circ+ genes also tend to be more expressed than expressed genes both in cells treated with scrambled or NELF-A siRNAs, although to a lesser extent than KD-only circ+ genes, with the exception of biological replicate 2 in cells treated with scrambled siRNA. These results suggest that the genes that respond to the siRNA treatment tend to be more expressed, consistent with our general observations in previous sections of the thesis, but that the genes for which circRNAs are detected in NELF-A knockdown tend to be more highly expressed than those detected only in scrambled siRNA treated cells. Although interesting, these results are very subtle and current work in our laboratory has expanded these analyses to two additional biological replicates.

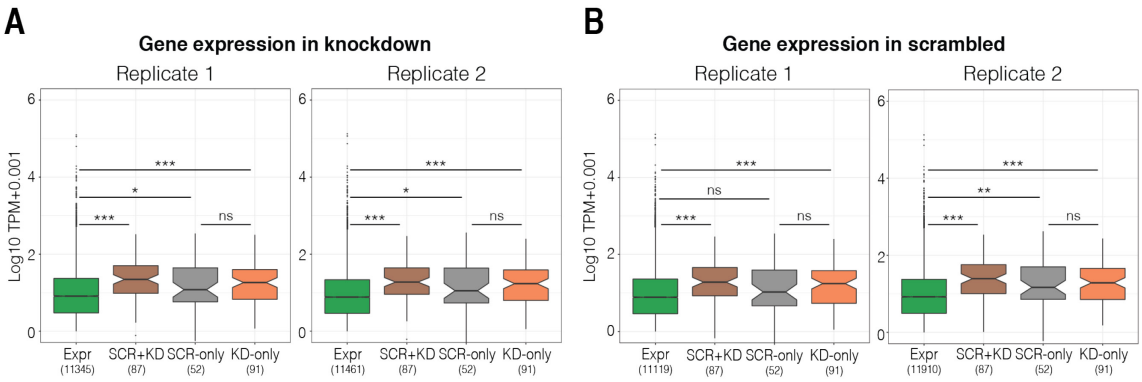

Figure 4.27 Expression of circRNA-producing genes.
Boxplots show the comparison of gene expression (TPM) in cells treated with **A)** scrambled siRNA or **B)** with NELF-A siRNA of genes producing circRNAs only in SCR (SCR), only in KD (KD) or both (SCR+KD) and expressed genes with TPM ≥ 1(Expr). Results are shown for replicates 1 and 2 separately. Wilcoxon rank sum test; * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001; ns – not significant.

4.8.6 Genes producing circRNAs upon NELF depletion do not have distinctive structural features

I next explored the structural features of genes that make circRNAs in NELF-A siRNA, scrambled siRNA or in both treatments, including gene length, exon and intron length, position of first and last included exon, and number of intervening exons. Similarly to previous results, all gene groups are composed of very long genes (> 25 kb) with many exons (>5) (**Fig. 4.28** and **Fig. 3.12**).

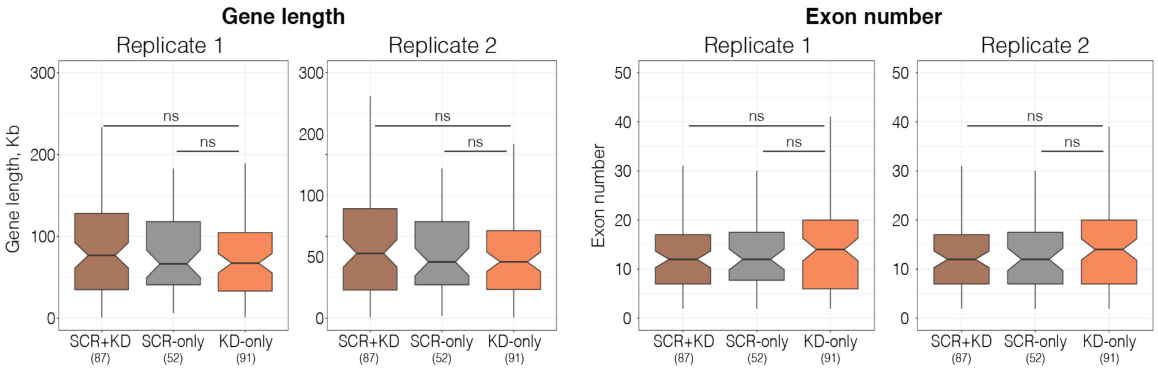

Figure 4.28 Structural features of circRNA producing genes in NELF-A KD experiments.
Boxplots show comparison of gene length (kb) and exon number of genes producing circRNAs only in SCR (SCR), only in KD (KD) or both (SCR+KD). Analyses are shown for individual biological replicates. SCR and KD gene isoforms correspond to each biological replicate. For SCR+KD, the mean value of transcript isoform from SCR and KD samples of each replicate was used. Wilcoxon rank sum test; ns – not significant.

I further determined the start and end position of circRNAs and how many exons are most often back-spliced. Similarly to published data and results shown in this work, circRNAs tend to start from the 5' end of genes, mostly at exon 2 (**Fig. 4.29A**), generally have 1 to 5 exons (**Fig. 4.29B**), and circRNAs tend to end close to the 5'end or middle of the gene for all gene groups (**Fig. 4.29C**).

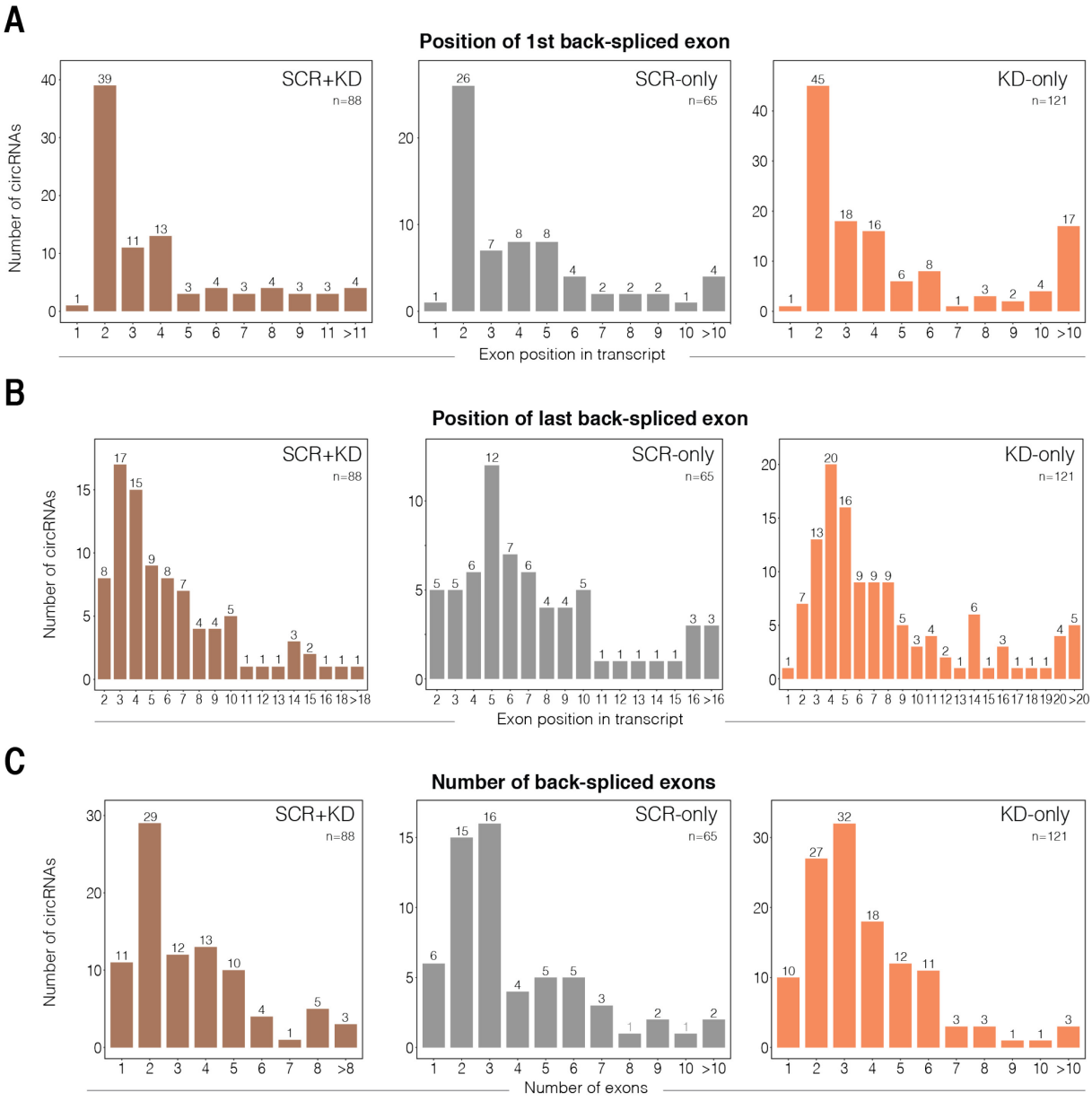

Figure 4.29 CircRNAs are produced from the 5' end of genes in NELF-A KD experiments.
A) Position of the first and **B)** last back-spliced exon in the transcript in circRNAs robustly detected only in SCR, KD or both. **C)** Number of exons included in circRNAs robustly detected only in SCR, KD or both. Analyses are shown for individual biological replicates. SCR and KD gene isoforms correspond to each biological replicate.

Finally, since circRNAs most often start at exon 2, I compared the length of intron 1 and exons 1 and 2 and, again, there were no differences between the three groups of genes (**Fig. 4.30**).

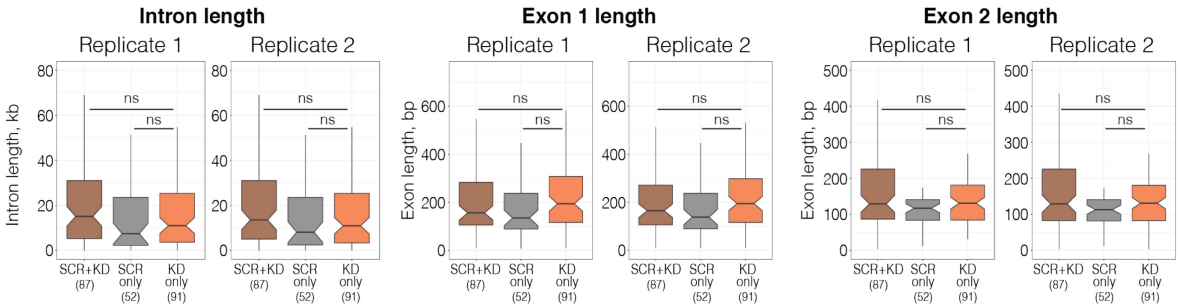

Figure 4.30 Length of the intron 1 and exons 1 and 2 of of circRNA-producing genes in NELF-A KD experiments.
Boxplots show intron 1 and exons 1 and 2 length for genes producing robustly detected circRNAs in SCR (SCR), only in KD (KD) or both (SCR+KD). Analyses are shown for individual biological replicates. SCR and KD gene isoforms correspond to each biological replicate. For SCR+KD, the mean value of transcript isoform from SCR and KD samples of each replicate was used. Wilcoxon rank sum test; ns – not significant.

No statistically significant differences were identified between genes making circRNAs in NELF-A or scrambled siRNA samples, and the new circRNAs identified here have the same properties previously found or untreated mESCs, neuronal progenitors and neurons (**Fig. 3.12** and **3.13**). Taken together, these analyses show that the detection of circRNAs in the NELF-A or scrambled siRNA samples originates from the same types of genes, suggesting that they all in general have a probability of generating circRNAs in ESCs.

4.8.7 Connecting the effects of NELF depletion with promoter dynamics in mESCs

Given that genes producing circRNAs are in general depleted for spliceosome, RNAPII modifications and promoter-proximal pausing modulators, I reasoned that KD-only genes could display increased sensitivity to changes in promoter-proximal pausing dynamics due to having different levels of these factors from non-transfected mESCs. To address this, I first compared the enrichment level at the TSS and E1-I1 border of RNAPII S7p, NELF and U1 snRNP at SCR+KD, SCR-only and KD-only genes in mESCs. Genes never producing circRNAs in mESCs (C-

ESC, filtered as in **Fig. 4.5**) and genes producing robust circRNAs in untreated mESCs (C+ ESC) are shown as comparison. Bulk analysis reveals that the distributions of enrichment levels of RNAPII S7p, NELF and U1 snRNP are similar across all the groups of genes considered both at the TSS (**Fig. 4.32, top panel**) and E1-I1 border (**Fig. 4.32, bottom panel**). Interestingly, all genes making circRNAs after transfection show higher levels of these markers in untreated mESCs, which may explain why they were not found to produce circRNAs in untreated cells, and only after siRNA treatment.

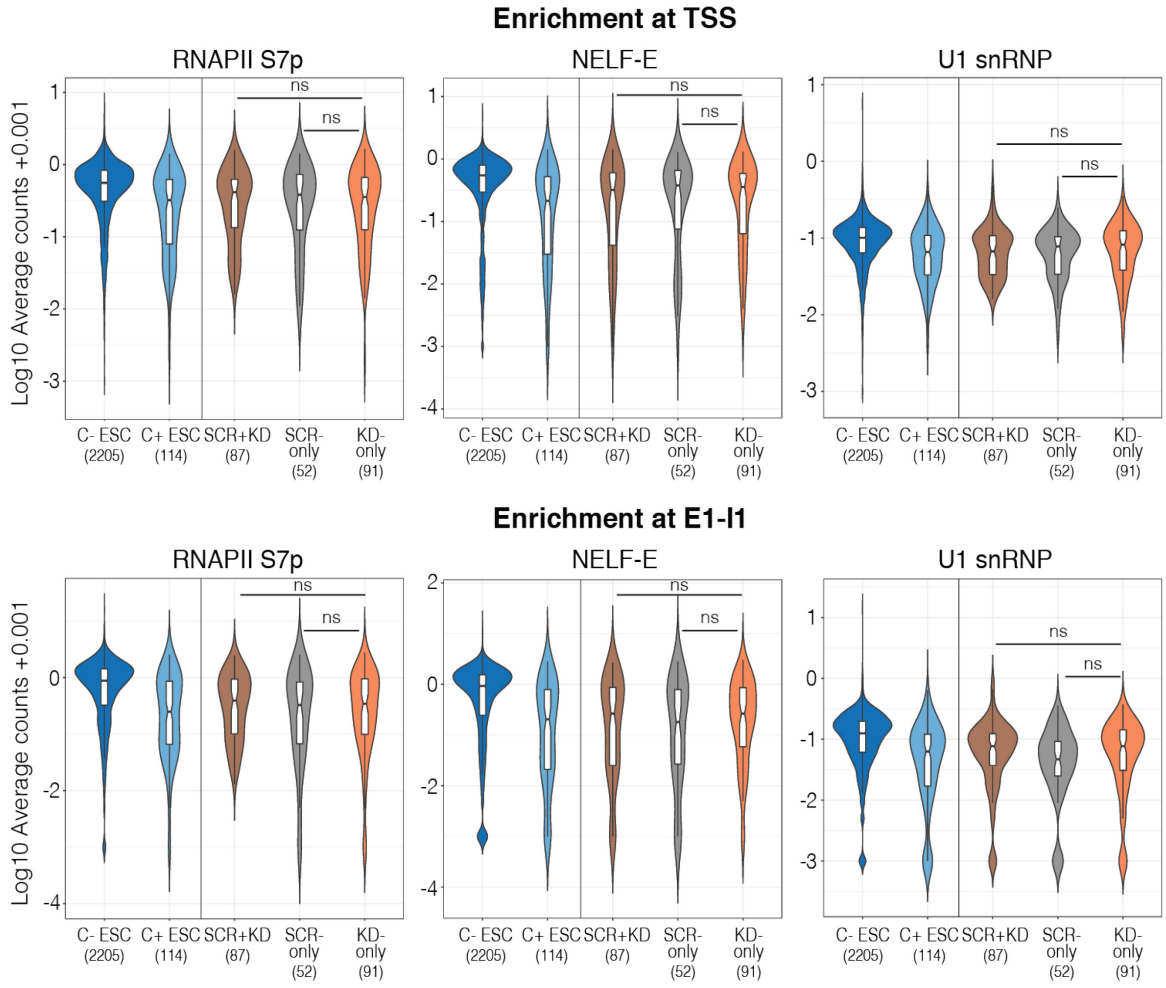

Figure 4.31 Enrichment of RNAPII S7p, NELF-E and U1C at genes producing robust circRNAs only in SCR, KD or common in both in mESCs.
Violin plots show enrichment of RNAPII S7p, NELF-E and U1 snRNP at TSS and E1-I1 border. Genes never producing circRNAs (C- ESC) and genes producing robust circRNAs in mESCs (C+ ESC) are shown as comparison. Wilcoxon rank sum test; ns – not significant.

To further dissect whether some genes within each group have specific features of RNAPII S7p, NELF and U1 snRNP enrichment, I divided all expressed genes in mESCs according to low, mid and high enrichment levels at the TSS of RNAPII S7p, NELF and U1 snRNP, with each group having ~ 3700 genes (**Fig. 4.33**).

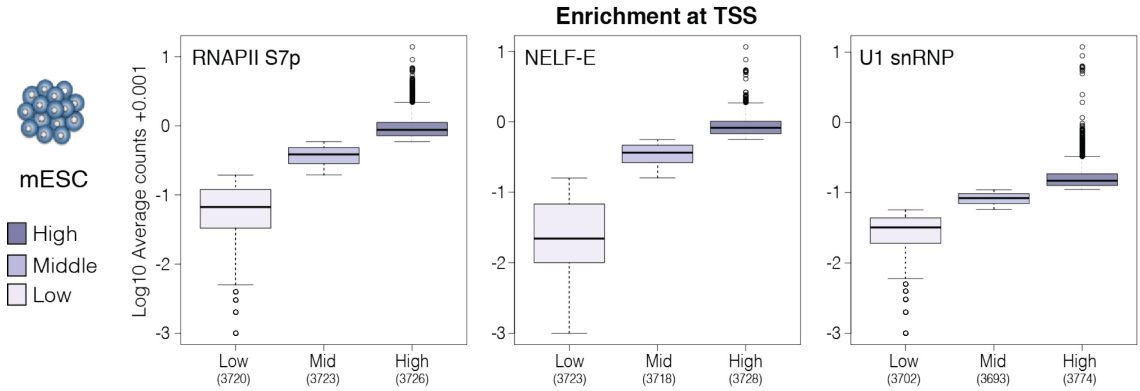

Figure 4.32 Quantile of enrichment levels for RNAPII S7p, NELF-E and U1 snRNP in mESCs.
Boxplots show enrichment level of RNAPII S7p, NELF-E and U1 snRNP. Expressed genes with TPM ≥ 1 in mESCs were divided into 3 quantiles corresponding to low, mid and high enrichment levels. Number of genes in each category is depicted on the x-axis.

I next quantified the percentage of SCR+KD, SCR-only and KD-only genes that overlap with each enrichment class. Genes which produce robust circRNAs in mESCs are shown as comparison. I started by analyzing the results corresponding to S7p enrichment. Confirming previous results, most circRNA producing genes in mESCs have low enrichment of RNAPII S7p, followed by mid and high enrichment (**Fig. 4.34, top row**). When looking at SCR+KD, SCR-only and KD-only genes, the proportion of genes in each category is comparable, where most genes show an intermediate enrichment level and fewer genes have high/low enrichment. When observing NELF enrichment levels in mESCs, there is an enrichment pattern similar to RNAPII S7p, with most genes having low levels of NELF, followed by mid and high levels and a similar pattern is observed for SCR+KD genes (**Fig. 4.34** middle panel). However, while most SCR-only genes have intermediate levels of NELF followed by low and high levels, KD-only genes show a similar proportion of genes for all enrichment categories, suggesting that KD-only genes may have slightly higher NELF levels. Finally, I compared U1 snRNP enrichment levels for all groups of genes producing circRNAs. In mESCs and SCR+KD genes, U1 snRNP

enrichment pattern is comparable to RNAPII S7p and NELF enrichment, with most genes having low levels (**Fig. 4.34** bottom panel). For SCR-only genes, most have low and intermediate levels, with very few genes having high levels of U1 snRNP. Interestingly, KD-only genes show a pattern distinct of SCR-only, with genes falling almost equally in low/mid/high categories, suggesting that KD-only genes may have a slightly higher level of U1 snRNP. The differences between all gene groups are very small due to the low number of genes used in these analyses and are therefore not significant. Nevertheless, when taken at face value, these results suggest that genes which produce circRNAs specifically in KD tend to have slightly higher levels of NELF and U1 snRNP, which could make them more sensitive to NELF-A depletion and interference with the promoter-proximal pausing mechanisms it regulates.

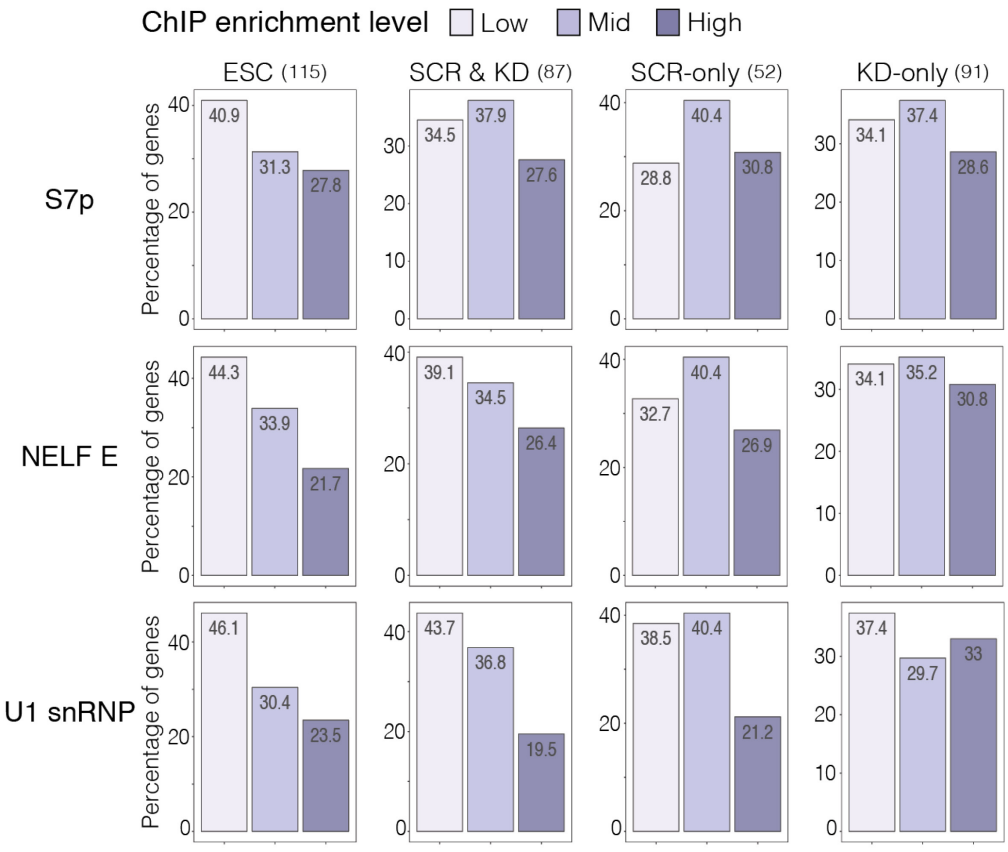

Figure 4.33 Percentage of genes with low, mid and high enrichment levels which produce circRNAs in mESCs, specifically in SCR, KD or both.
Percentage of circRNA-producing genes per quantile of ChIP-seq enrichment level at the TSS for RNAPII S7p, NELF-E and U1 snRNP. Number of genes in each gene group is shown on top and percentage values are represented on top of each bar.

4.9 Discussion

CircRNA formation is a very complex process and evidence suggests that circRNA biogenesis may be associated with co-transcriptional splicing regulation: genes producing circRNAs are transcribed by a faster elongating RNAPII (Ashwal-Fluss et al. 2014; Zhang et al. 2016) and, in some instances, decreasing splicing or polyadenylation efficiency by knocking down spliceosome components or cleavage and polyadenylation factors increases circRNA production (Liang et al. 2017). Additionally, mechanics of the back-splicing reaction imply that at least a temporary decoupling between transcription and splicing is necessary to form circRNAs. Thus, circRNA formation may hinge on spliceosome recruitment to nascent RNA molecules, which in turn could depend on RNAPII post-translational modifications, for example S5p and S2p. However, it is not known how interplay between RNAPII modifications and spliceosome recruitment influences circRNA formation. Moreover, it is unclear how transcription-splicing dynamics influence circRNA formation in different cell types. To answer these questions, I explored in depth the relationship between circRNA biogenesis, spliceosome recruitment, and RNAPII post-translational modifications in mESCs and throughout neuronal maturation.

4.9.1 Genes producing circRNAs have altered spliceosome recruitment, RNAPII S5p and S7p levels and promoter-proximal pausing dynamics in mESCs

After designing a strategy to study genes producing circRNAs, I first investigated the interplay between spliceosome recruitment and RNAPII modifications at these genes in mESCs. I started my explorations in this cellular system because it is easier to manipulate experimentally, it is very well characterized, and with a large resource of ChIP-seq datasets published in this system. I first mapped the occupancy of U1 snRNP on chromatin, which demonstrated that circ+ genes are depleted for U1 snRNP at the promoter and exon1-intron1 junction when compared to circ- genes. RNAPII S5p, an essential RNAPII modification for proper co-transcriptional recruitment of the spliceosome is also depleted at circ+ genes

compared with genes that were not found to make circRNAs in the seven stages of differentiation considered here. Remarkably, S7p was also strongly depleted at the promoters of circ+ genes, while S2p levels are unchanged at its most abundance regions of enrichment, which are immediately upstream or downstream of the polyadenylation site. These results indicate that the transition from initiation into early stages of elongation is altered at circ+ genes, but that RNAPII elongates and terminates transcription properly. High mRNA expression levels from circ+ genes support these findings. Interestingly, total RNAPII is slightly decreased at the TSS of circ+ genes, but not to the same extent as S5p and S7p. Thus, the lower levels of S5p and S7p at circ+ genes are not fully explained by reduced RNAPII recruitment.

I next asked which stages of the transcription cycle were altered at circ+ genes: RNAPII recruitment, transcription initiation or RNAPII release into productive elongation, i.e. promoter-proximal pausing. While RNAPII recruitment and transcription initiation are mostly unaffected, promoter-proximal pausing factors NELF and CDK9 are markedly depleted at circ+ genes.

It seemed counterintuitive that circ+ genes are highly expressed while at the same time being depleted of S5p, S7p, NELF and CDK9. Accumulation of ChIP-seq signal at the promoter reflects the dynamics of recruitment, binding, and release of protein complexes within the cell population captured upon fixation (Ehrensberger, Kelly, and Svejstrup 2013). At circ+ genes, the main contributors to decreased S5p and S7p enrichment seem to be promoter-proximal pausing factors (NELF and CDK9). Decreased NELF and CDK9 enrichment together with high levels of gene expression indicate that RNAPII is recruited to circRNA-producing genes, being quickly released into productive elongation (possibly without the appropriate modifications). Decreased CDK9 levels likely reflect reduced levels of NELF, since S2p is not altered at circRNA-producing genes. The premature release of RNAPII from pausing into productive elongation may lead to sub-optimal recruitment of U1 snRNP (and possibly of other spliceosome subunits), ultimately favoring circRNA formation. It is worth noting that the

analyses conducted in this work do not address the possibility that high expression levels of circ+ genes could result from increased mRNA stability. This could be clarified by, for example, quantifying the level of nascent RNA produced at genes producing circRNAs.

4.9.2 Genes producing circRNAs in neurons also show altered promoter-proximal pausing dynamics in neurons

CircRNAs are far more abundant in neurons (Rybak-Wolf et al. 2015; You et al. 2015; Zhang et al. 2016), so I set out to study transcription-splicing dynamics at circRNA-producing genes in neurons. I found that RNAPII S5p and S7p are slightly decreased at these genes in dopaminergic neurons days 16 and 30. The milder depletion of S5p and S7p in dopaminergic neurons could suggest increased RNAPII recruitment at circ+ genes, causing RNAPII to accumulate to higher levels at the TSS or a different regulation of pausing in these cells. It was not possible to immunoprecipitate U1 snRNP in dopaminergic neurons, preventing our ability to confirm whether U1 snRNP recruitment is altered at circ+ genes in dopaminergic neurons. This could be explained by the epitope for U1C antibody being masked or by U1 snRNP binding less to RNAPII complexes in dopaminergic neurons and thus being more difficult to immunoprecipitate. One possible solution could be performing ChIP for U1 snRNP in the presence of RNAse inhibitors to help preserve its binding to RNAPII complexes. Nevertheless, similarly to mESCs, NELF is markedly depleted at circ+ genes, which suggests that a mechanism analogous to mESCs underlies circRNA formation in dopaminergic neurons.

Finally, RNAPII S5p is also decreased at circ+ genes in spinal motor neurons days 2 and 8. This effect seems more prominent than in dopaminergic neurons and comparable to mESCs, possibly due different regulatory mechanisms in both differentiation systems. Given the marked depletion of S5p at circ+ genes in spinal motor neurons, it is likely that RNAPII promoter release and spliceosome recruitment are also altered in these cells. To confirm this, it would be necessary to perform ChIP of RNAPII S7p, NELF-E, and U1C in spinal motor neurons.

Altogether, results suggest that release of RNAPII from the promoter and (probably) spliceosome recruitment may contribute to circRNA formation in very different cell types. It would be very interesting to extend these observations to a non-neuronal system where circRNA formation also occurs quite frequently, for example in cardiac cells.

4.9.3 Depletion of NELF increases the likelihood of circRNA production in mESCs

Studies performed in mESCs and neurons strongly suggest that increased release of RNAPII from the promoter leads to altered spliceosome recruitment and favors circRNA production. If this is the case, perturbing promoter-proximal pausing should impact circRNA biogenesis. To test this, I knocked down NELF-A subunit to trigger RNAPII release from the promoter and test whether deregulation promoter-proximal pausing consequently increased circRNA production. Firstly, we find that NELF knockdown does not impact gene expression, which is counterintuitive since promoter-proximal pausing should be affected. Published literature did not evaluate the effects of NELF knockdown on gene expression in mESCs and a direct comparison is lacking. Most papers where NELF was depleted base their findings on gene expression arrays and report a broad range of differentially expressed genes, from ~100 to 500 genes (Gilchrist et al. 2008), (Narita et al. 2007), (Sun and Li 2010), where most genes are downregulated. Another study evaluated the effects of NELF knockdown by poly(A) selected RNA-seq and reports ~ 2700 downregulated, with very few up-regulated genes (Stadelmayer et al. 2014). The discrepancies between our study and others could be explained by interfering with different NELF subunits, different methods used to quantify gene expression and/or different cell lines studied.

Next, I evaluated the effect of NELF knockdown on circRNA production. CircRNA detection is highly variable between biological replicates for both scrambled and NELF-A knockdown samples. High variability of circRNA detection between biological replicates could indicate that many circRNAs are very lowly expressed and very close to the detection threshold of current technologies. It is likely that

every time a sample is produced for circRNA detection, some circRNAs are sampled from the pool of lowly expressed circRNAs, causing high variability between biological replicates. For example, although we find many circRNAs common in scrambled and NELF knockdown samples, others are only detected in scrambled or knockdown samples. Some circRNAs are detected only in scrambled are also found in mESCs, suggesting that these circRNAs could already be present in the cell. However, others are not found in mESCs and could either result from spurious transfection effects or be part of the "lowly-expressed circRNAs" which are randomly sampled in each biological replicate. CircRNAs detected only in knockdown samples are of course subject to a similar bias. We find that most genes producing circRNAs in scrambled and/or knockdown samples also produce robustly detected or uniquely detected circRNAs at least once during neuronal differentiation, especially at days 16 and 30 of dopaminergic neuron differentiation, where circRNAs tend to accumulate, suggesting that these genes have the potential to form circRNAs. Nevertheless, when taken at face value, we do detect more circRNAs and genes producing circRNAs upon NELF knockdown when compared to the scrambled sample. This trend is evident in independent biological replicates, suggesting that NELF knockdown does increase the likelihood of circRNA formation and is sufficient to induce circRNA formation. At the same time, we do not observe changes in the expression of circRNAs common to scrambled and knockdown samples. This could be due to genes producing these circRNAs being less sensitive to fluctuations in NELF levels, as NELF would already be slightly decreased at many of these genes. Finally, we cannot rule out that the effects of NELF knockdown on circRNA production are indirectly due to defects in snRNP function. Studies have shown that NELF knockdown causes defects in snRNA processing and is therefore important for proper expression of snRNAs (Egloff et al. 2009; Yamamoto et al. 2014). Although most proteins that are part of snRNPs were not affected, we did not evaluate the effects of NELF knockdown on snRNAs, which could indirectly alter spliceosome function and increase circRNA formation.

I further explored the features of genes producing circRNAs only in scrambled, knockdown or both. Results showed that genes within these groups are more highly expressed than average, without significant differences among them. This analysis also revealed that even when genes do not produce circRNAs, they still tend to be highly expressed, suggesting that high expression level is not a pre-requisite for circRNA formation. Moreover, genes producing circRNAs only in scrambled, knockdown or both are structurally similar: these are very long and have many exons, produce circRNAs most often from exon 2 and showed comparable length of intron 1 and exons 1 and 2.

I finally explored the enrichment levels of RNAPII S7p, NELF and U1 snRNP in mESCs for genes producing circRNAs only in scrambled, knockdown or both. Bulk analyses did not show any differences between the three groups. I then categorized genes from the three groups according to low, mid and high enrichment levels. Although not statistically significant, genes producing circRNAs only in knockdown tend to have slightly higher levels of NELF and U1 snRNP when compared to genes producing circRNAs only in scrambled, both scrambled and knockdown or mESCs, suggesting that genes which start producing circRNAs upon NELF knockdown may be more sensitive to fluctuations of NELF levels and rely more on RNAPII pausing to timely recruit the spliceosome to the nascent RNA. However, we do not know how NELF knockdown affects RNAPII modifications and spliceosome recruitment with the current analysis. To further clarify this, it would be necessary to map the occupancy on chromatin of RNAPII S5p/S7p and U1 snRNP in NELF-depleted mESCs.

Altogether, NELF knockdown effects in circRNA production are noticeable but relatively mild. In line with this, Rahl and colleagues reported that only some genes showed shifted RNAPII enrichment from the promoter into the gene body upon NELF-A knockdown (Rahl et al. 2010). It would be interesting to determine whether the knockdown of a different NELF subunit, for example the catalytic subunit NELF-E, would cause stronger effects in circRNA production. It would

also be very interesting to study the effects when depleting DSIF or NELF and DSIF complexes simultaneously, as these two complexes may compensate each other's function. An alternative and more targeted approach to studying the mechanics of promoter-proximal pausing at genes producing circRNAs in mESCs, would be fusing CDK9 to dead Cas9 and bringing it to the promoter of candidate genes to trigger RNAPII promoter release and increase circRNA formation. This is currently being done in collaboration with Stefan Stricker's group (Institute of Stem Cell Research, Munich, Germany), but so far did not yield clear results and requires further optimization (not shown). Finally, although technically very challenging, it would be most interesting to devise strategies that can disrupt promoter-proximal pausing mechanisms in neurons, where circRNAs are most abundant. This could be achieved, for example, through the generation of stable mESCs cell lines where NELF or DSIF subunits would be fused with auxin-inducible degron (AID) system. In this way, degradation of NELF/ DSIF could be specifically induced in dopaminergic neurons, without affecting cell differentiation.

To conclude, in this chapter I have shown that genes producing circRNAs have decreased U1 snRNP recruitment, which is likely due to decreased RNAPII S5p enrichment at the promoter of these genes. Furthermore, these genes show decreased RNAPII S7p together with lower levels of promoter-proximal pausing factors NELF and CDK9, pointing towards sub-optimal RNAPII pausing at the promoter and quick release into productive elongation. This mechanism appears to underlie circRNA formation not only in mESCs but also in dopaminergic neurons and spinal motor neurons. Finally, I have shown that depletion of NELF-A is sufficient to increase circRNA production in mESCs and that these genes tend to have slightly higher levels of NELF and U1 snRNP, showing that RNAPII release from the promoter does impact circRNA formation.

Part IV

Discussion

5 Discussion

CircRNAs were at first thought to be rare by-products of splicing due to their sporadic detection and to the idea that most splicing occurs co-transcriptionally and that introns are spliced in a "first come first served" manner. With the onset of next generation sequencing technologies, the discovery of large numbers of circRNAs in most cell types and in a vast array of organisms, ranging from yeast to humans, shows that they are a widespread class of RNAs (Li, Yang, and Chen 2018; Wilusz 2018). CircRNAs are most abundant in neuronal cells and tissues, especially enriched in synapses, suggesting that they modulate neuronal function. Growing evidence also supports a role of circRNAs in innate immunity and cancer (Chen, Satpathy, and Chang 2017; Patop and Kadener 2018). Understanding the mechanisms of circRNA formation has therefore become an increasingly important question in the field of RNA biology.

CircRNA formation is regulated at multiple levels, from the presence of complementary repeats in the introns flanking back-spliced exons, to the binding of RBPs that facilitate or prevent circRNA formation (Chen 2016; Wilusz 2018). Recent reports explored how circRNA formation depends on the dynamics between transcription and splicing. In both *Drosophila* and humans, the use of slow/fast elongating RNAPII mutants indicates that circRNA-producing genes are transcribed by a fast elongating RNAPII (Ashwal-Fluss et al. 2014; Zhang et al. 2016). Furthermore, reducing the efficiency of splicing or transcription termination, in *Drosophila*, led to the upregulation of some circRNAs (Liang et al. 2017). Nevertheless, the mechanisms by which transcription and splicing dynamics regulate circRNA formation are still unknown. It is also unclear how changes in splicing and transcription dynamics in specific genes and in distinct cell types contribute to the cell-type specific production of circRNAs.

To study how transcription and splicing dynamics relate with circRNA production, I explored transcriptome sequencing together with genome-wide

occupancy of the spliceosome complex, RNAPII post-translational modifications and transcription modulators. I took advantage of two time series of differentiation from mESCs into two types of neurons, dopaminergic and spinal motor, to investigate how the expression of circRNAs from specific genes and at specific times of differentiation relates with the RNAPII modification and the dynamics of spliceosome recruitment.

Genes producing circRNAs are highly expressed and produce circRNAs most often from exon 2

In chapter 3, I found that circRNAs are produced in mESCs and at all time-points of dopaminergic and spinal motor neuron differentiation, being most abundant in dopaminergic neurons day 16. A slightly lower detection at day 30 than day 16 was surprising since the literature shows that circRNAs tend to accumulate during neuronal development and aging (Gruner et al. 2016; Westholm et al. 2014; Rybak-Wolf et al. 2015; Zhang et al. 2016). Further investigation would be required to understand if decreased number of circRNAs in fully differentiated neurons is a specific feature of dopaminergic neurons, or a property of the in vitro differentiation used. I also found that circRNA expression is highly dynamic with 445 out of 881 circRNA-producing genes detected being time point specific (**Fig. 3.9**).

Previous genome-wide analyses showed that detection of back-splicing at specific genes correlates with decreased splicing of the intron upstream of the first circRNA exon (Ashwal-Fluss et al. 2014; Zhang et al. 2014). These observations provide an important test to whether the detection of circRNAs is specific and not an artefact of sequencing or bioinformatics analyses. However, these observations also lead to the idea that the linear mRNA is expressed less when the circRNAs are produced. I took advantage of the availability of RNA-seq datasets for the dopaminergic (total and poly(A) RNA-seq) and spinal motor neuron differentiation (total RNA-seq), and asked whether the increased detection of circRNAs at any given time point of differentiation related with lower level of expression of the linear transcript. Surprisingly, by quantifying mRNA

expression from both total and poly(A) RNA-seq data, I found that genes producing circRNAs are most often expressed in all the stages of differentiation investigated, although to varying levels. Unexpectedly, I found that production of circRNAs was increased when genes were more expressed, and that genes which produce circRNAs tend to be more highly expressed than genes that never produce circRNAs. The genome-wide analyses presented in chapter 3 suggest that circRNA production is associated with high transcription rate.

I also characterize the structural properties of the genes that make circRNAs. Several studies had previously shown that circRNAs usually have 1-5 exons, are often flanked by very long introns and are mostly produced from the 5' end of genes (Gruner et al. 2016; Jeck et al. 2013; Salzman et al. 2012). In line with these findings, circRNAs detected in the differentiation systems studied in this thesis also tend to have 1-5 exons, and to be produced from long genes with many exons. I further confirmed that circRNAs tend to be produced from the 5' end of genes with a strong bias for exon 2, where the first splicing reaction occurs. This together with increased transcription rate, indicates that not only the interplay between transcription and splicing could be altered at circRNA-producing genes, but also that promoter-based mechanisms could modulate circRNA production. These results suggest that the in vitro system chosen is a good model to investigate the molecular mechanisms that regulate circRNA biogenesis, and especially its relation with transcription and splicing dynamics.

The promoter regions of circRNA-producing genes have reduced spliceosome recruitment and are depleted for S5p and S7p modifications of RNAPII

In chapter 4, I explored the role of the interplay between transcription and splicing on circRNA formation. To investigate promoter-based mechanisms that could influence circRNA formation by impairing the recognition of the first exon-intron junction, I focused my efforts on exploring at the promoters and exon1-intron1 junctions. I started by dissecting the promoter-based mechanisms of circRNA biogenesis in mESCs, for which there is a large resource of published

genome-wide chromatin occupancy datasets, and for which it was easiest to produce new datasets.

U1 snRNP is the first spliceosome subunit recruited to splice sites and is essential for appropriate splice site recognition (Matera and Wang 2014; Will and Luhrmann 2011). To investigate whether circRNA genes have decreased U1 snRNP recruitment at their promoters, I performed by ChIP-seq of the U1C protein. I found that genes producing circRNAs are depleted for U1 snRNP at the promoter and especially at the exon1-intron1 junction (**Fig. 4.6**), indicating that spliceosome recruitment is altered at circRNA-producing genes.

Although RNAPII recruitment (TBP and TAF1) and transcription initiation (CDK7 and 8) is similar at circRNA-producing and non-producing genes, S5p is markedly depleted at the former genes. RNAPII S5p was shown to interact with U1 snRNP, and with other spliceosome subunits, and is thought to play an important role in spliceosome recruitment during co-transcriptional RNA processing (Harlen et al. 2016; Nojima et al. 2018). As such, lower levels of S5p at circRNA-producing genes (**Fig. 4.7**) is consistent with spliceosome recruitment.

Furthermore, I found that S7p is depleted at the promoters of circRNA-producing genes. S7p facilitates the transition from RNAPII initiation to elongation, by priming the RNAPII CTD for PTEF-b recognition and phosphorylation of S2 by CDK9 (Czudnochowski, Bosken, and Geyer 2012; St Amour et al. 2012; Viladevall et al. 2009). As such, reduced S7p levels could imply deficient modification of RNAPII or impaired promoter-proximal pausing. Finally, the abundance of S2p is unchanged at the TES of circRNA-producing genes, consistent with these genes being actively expressed at the mRNA level. This suggests that elongation and termination occur efficiently at circRNA-producing genes and reinforces that promoter-proximal pausing could be affected.

RNAPII release from promoter-proximal pausing is altered at circRNA-producing genes in mESCs and neurons

To understand whether promoter-proximal mechanisms are altered, I investigated the abundance of NELF and CDK9, both modulators of RNAPII promoter-proximal pausing, and found that they are markedly depleted. This, together with unchanged S2p and high gene expression levels, suggests that RNAPII is recruited and phosphorylated upon transcription initiation, but that it is quickly released into the gene body. RNAPII reduced time at the promoter-proximal pausing site, possibly without complete modifications of the CTD (S5p and S7p), may lead to decreased spliceosome recruitment, hence favoring circRNA formation instead of canonical splicing.

An interesting possibility is that promoter-proximal pausing factors may also impact on the recruitment of the splicing machinery for the timely recognition of the first exon-intron junction. For example, a recent study in yeast showed that SPT5 interacts with all snRNPs and that degradation of SPT5 by AID reduces U5 snRNP recruitment to intron-containing genes, suggesting that SPT5 helps recruiting snRNPs during co-transcriptional splicing (Maudlin and Beggs 2019). So far, little evidence supports a direct role of NELF in splicing regulation. NELF was shown to help processing the 3'end of histone mRNAs together with CBC in HeLa cells (Narita et al. 2007). A proteomic study in HeLa cells identified NELF-E bound together with U1/U2 snRNPs using a gene construct where the commitment complex formed (Sharma et al. 2008). Nevertheless, this was a highly artificial system and based on a single gene. Thus, it remains to be understood whether NELF could modulates splicing directly.

As circRNAs are most abundant in neurons, it was important to verify whether the major observations made using mESCs were also true in neurons. Therefore, I next investigated whether RNAPII S5p and NELF were also depleted at the promoters and exon1-intron1 junctions in dopaminergic and spinal motor neurons. After producing new datasets and exploring publicly available datasets, I found similar distributions to mESCs which suggests that RNAPII release from the

promoter is also associated with circRNA formation in dopaminergic neurons. Similarly to mESCs and dopaminergic neurons, RNAPII S5p is depleted at genes producing circRNAs in spinal motor neurons. To further verify if promoter-proximal pausing is altered in these cells, it would be important to map the occupancy of NELF and U1 snRNP in spinal motor neurons. Interestingly, in dopaminergic neurons, although NELF is markedly depleted at the TSS of circRNA-producing genes, RNAPII S5p and S7p depletion is not as profound as NELF. It is possible that RNAPII recruitment, transcription initiation, or promoter-proximal pausing are regulated slightly differently in dopaminergic neurons. It would be very interesting to explore transcription dynamics in these cells by, for example, mapping the occupancy of GTFs and CDK7/8, to understand how these complexes modulate transcription in dopaminergic neurons. Finally, given the depletion of S5p, S7p and NELF, it is likely that U1 snRNP recruitment is altered in these dopaminergic neurons. I attempted to address this with ChIP of U1 snRNP, but could not obtain a specific enrichment. Further optimizations are necessary in the future to map U1 snRNP occupancy on chromatin by ChIP in dopaminergic neurons.

NELF depletion is sufficient to induce circRNA production

In the last part of chapter 4, I aimed to mechanistically test the role of Promoter-proximal pausing in circRNA formation, by interfering with the levels of NELF, a negative regulator of elongation, which contributes to the slowing down of RNAPII before elongation starts. We hypothesized that by perturbing RNAPII pausing genome-wide and and promoting its release into productive elongation would reduce spliceosome recruitment and increase circRNA formation in a larger number of genes. To disrupt RNAPII pausing, I performed RNA interference experiments to knockdown NELF-A. NELF knockdown was efficient and did not have detectable effects in mRNA production genome-wide. When investigating the synthesis of circRNAs, we detected a higher number of circRNAs and of circRNA-producing genes in NELF knockdown experiments compared to control treatment using scrambled siRNAs. These results indicate that disruption

of RNAPII pausing at gene promoters is sufficient to promote circRNA formation. In the future, it would be of great interest to further dissect the mechanisms by which NELF depletion leads to increased circRNA production, namely by studying RNAPII modifications and U1 snRNP recruitment.

5.1 A model for circRNA biogenesis

Based on the results shown in this thesis, I propose a model for circRNA formation in both mESCs and neurons. RNAPII is recruited to gene promoters and initiates transcription. At circRNA-producing genes, altered promoter-proximal pausing that coincides with reduced levels of NELF, leads to a quick release of RNAPII from the promoter into early elongation without the sufficient levels of CTD modifications (S5p and S7p). This in turn leads to reduced spliceosome recruitment to the nascent RNA, which likely causes a temporary decoupling between transcription and splicing and ultimately favoring circRNA production (**Fig. 5.1**). It is likely that this model synergizes with other factors that modulate circRNA formation, such as inverted complementary repeats or RBPs, which act on the unprocessed transcript to achieve optimal circRNA expression in different cell types.

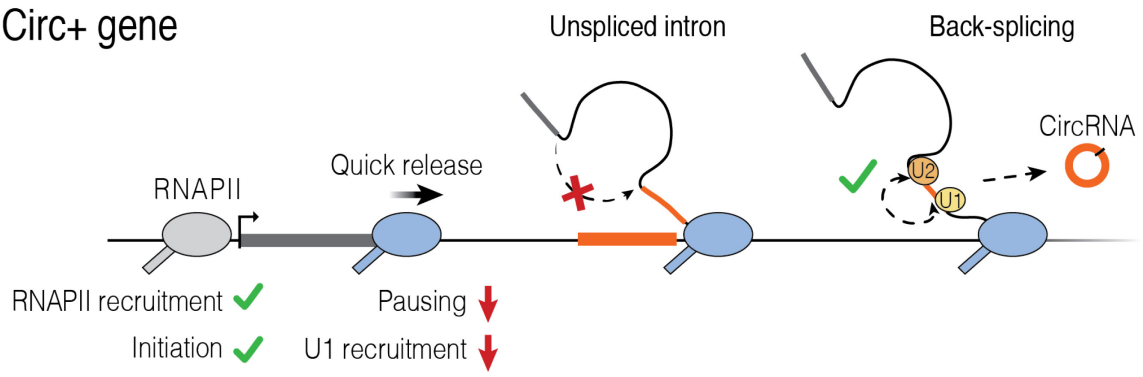

Figure 5.1 Model for circRNA formation.
RNAPII is recruited to the promoter of genes producing circRNAs (Circ+ gene) and initiates transcription. Reduced promoter-proximal pausing at circRNA-producing genes causes RNAPII quick release intro productive elongation, without the appropriate modifications. This in turn reduces spliceosome recruitment to nascent RNA, favouring back-splicing and circRNA formation.

We wish to further test the proposed model. For example, if specific promoter features (e.g. transcription modulators, epigenetic features, CG content, etc) are found to be strongly associated to circRNA production, we would be able to further dissect the mechanisms that lead to lower NELF recruitment and help predict which genes produce circRNAs in different cell types. We are currently investigating features that distinguish the promoters of circ+ and circ- genes, in collaboration with Prof. Martin Vingron. Firstly, we are comparing genes producing circRNAs with genes not producing circRNAs in mESCs to determine which promoter features are more predictive of circRNA formation. Based on these features, we next wish to predict circRNA formation in dopaminergic neurons.

Recruitment versus kinetic models at circRNA-producing genes

Genes producing circRNAs are transcribed by a fast elongating RNAPII (Ashwal-Fluss et al. 2014; Zhang et al. 2016). Importantly, the model for circRNA production proposed in this thesis does not conflict with this finding and it is likely that both RNAPII release from pausing with altered spliceosome recruitment and fast elongation speed at the gene body contribute to circRNA formation. I argue that one or the other mechanism may take precedence depending whether RNAPII is transcribing close to the promoter or further along the gene body. For example, RNAPII elongation speed was shown to accelerate as transcription progresses through the gene body (Jonkers, Kwak, and Lis 2014). Moreover, the Caceres group explored the effect of RNAPII elongation speed on alternative splicing in mESCs and neurons using mutant knock-in cells of a slow RNAPII and showed that alternative splicing was affected, especially at long genes in NPCs and neurons (Maslon et al. 2019). Interestingly, the speed of RNAPII elongation appears comparable between wildtype and slow mutant at the first exon-intron junction in at least two genes tested. Thus, the literature suggests that the contribution of RNAPII elongation speed becomes more relevant for splicing regulation as RNAPII progresses through the gene body. The experiments described in this thesis did not explore whether RNAPII elongation speed is

important for circRNA formation close to the promoter. Addressing this question is technically challenging, as RNAPII elongation speed is calculated based on newly synthesized RNA which is transcribed after RNAPII release from DRB inhibition. . We have obtained the mutant knock-in cells of a slow RNAPII from the group of Javier Caceres. Dr. Alexander Kukalev and Dr. Mohamad Aboelenin (Pombo lab) have produced total RNA-seq datasets, and the circRNA analyses are currently being performed by Petar Glažar from the group of Nikolaus Rajewsky to further dissect the contribution of RNAPII speed in circRNA formation.

5.2 Promoter-proximal pausing modulates circRNA production – future directions

The most exciting outcome of this thesis is that RNAPII release from the promoter modulates circRNA formation. This work unveils a network of interactions where the dynamics of promoter-proximal pausing, mediated by RNAPII modifications, modulates spliceosome recruitment and impacts the outcome of splicing. Two important questions that naturally stem from this thesis are: is circRNA production a side-effect of increased mRNA expression, or is promoter-proximal pausing specifically regulated to modulate the production of specific circRNAs, while ensuring the synthesis of the linear RNAs?

Potential mechanisms for modulating circRNA biogenesis through RNAPII promoter release

Results from this thesis and the literature show that circRNA expression is mostly cell-type specific (Rybak-Wolf et al. 2015; Salzman et al. 2013). Considering the proposed model, circRNA formation could be controlled by modulating the rate of RNAPII release from pausing. This could in turn be regulated in many ways, for example, through the recruitment rate of RNAPII mediated by transcription factors and chromatin remodelers and/or the expression levels of promoter-proximal pausing factors. Recent evidence indicates that RNAPII pausing at promoters, rather than RNAPII recruitment, is the limiting step in the transcription cycle (Bartman et al. 2019; Gressel et al. 2017; Shao and Zeitlinger

2017). Hence, it is expected that this step is intricately regulated by different complexes and in different ways.

Promoter-proximal pausing could be modulated through the recruitment of NELF and/or SPT5 to promoters. Although not much is known, some studies shed light on how NELF could be recruited to genes. In *Drosophila*, GAGA factor (GAF) brings NELF to the promoter of genes before transcription initiates and loads it onto RNAPII (Li et al. 2013). However, this mechanism has so far not been described in mammalian cells. NELF also interacts with c-fos and c-jun proteins, suggesting that some transcription factors could recruit NELF to genes (Aiyar et al. 2004; Zhong et al. 2004). Interestingly, CTCF was shown to promote the recruitment of NELF, SPT5 and CDK9 at *c-myc* gene, raising the possibility that promoter-proximal pausing factors could be recruited by CTCF to genes whose expression depends on CTCF binding near promoters and that genome architecture could influence RNAPII pausing (Laitem et al. 2015).

Several complexes which modulate promoter-proximal pausing are also known to regulate or be regulated by promoter-enhancer contacts. For example, PAF1, which modulates promoter-proximal pausing, also mediates promoter-enhancer contacts (Chen et al. 2015; Chen et al. 2017; Yu et al. 2015). The mediator subunit MED26, present at enhancers, contributes to the recruitment of the SEC to promoters, which in turn triggers RNAPII release via PTEF-b (Lens et al. 2017; Takahashi et al. 2011). Importantly, BRD4, which also controls RNAPII promoter release, is not only a scaffold protein that recruits chromatin modifiers, transcription factors and super enhancers to genes (Dey et al. 2000; Floyd et al. 2013; Loven et al. 2013), but also itself modifies the chromatin landscape through histone acetylation and nucleosome eviction (Devaiah et al. 2016). This raises the possibility that circRNA production could, directly or indirectly, be regulated by the chromatin environment. Indeed, we are exploring this in collaboration with Martin Vingron group to understand why transcription regulation is different in genes producing circRNAs. We are currently using numerous ChIP-seq published datasets in mESCs to interrogate the interaction networks between transcription

modulators, transcription factors, histone modifications and modifiers at the promoters of genes producing circRNAs and genes not producing circRNAs.

Biological impact of connecting circRNA formation to promoter-proximal pausing

One conclusion from this work is that the recruitment of processing machineries to the nascent RNA at early stages of transcription is essential for co-transcriptional splicing which, when altered, can lead to the production of circRNAs. It is possible that RNAPII release from the promoter could also modulate alternative splicing through the differential recruitment of splicing modulators. This view is supported by reports where the type of promoter and transcriptional activators can influence alternative splicing through the differential recruitment of splicing modulators (Cramer et al. 1999; Cramer et al. 1997; Kadener et al. 2001; Nogues et al. 2002). This possibility can be explored, for example, by studying alternative splicing in cells where promoter-proximal pausing mechanisms were disrupted.

Given that most circRNAs detected in this, and other studies, are very lowly expressed compared to their corresponding linear transcripts and that circRNAs are made from highly expressed genes, could circRNAs be a by-product of splicing reactions? One possible answer is yes, they could be. Another interesting possibility is that some circRNAs could be produced to fine-tune expression of the linear transcripts upon cellular stimuli, when the linear transcript produced after back-splicing is degraded. If highly expressed genes are consistently active and transcribing, perhaps it would be more energetically advantageous for the cell to trigger circRNA formation via RNAPII release from the promoter, rather than shutting down transcription altogether. Another exciting possibility is that circRNAs could be used to indirectly modulate protein isoforms produced within the cell. For this to happen, the linear transcript would not be degraded after back-splicing, but instead translated into a functional truncated protein where the domains encoded by circularized exons would be missing. A study by Kelly and colleagues supports this hypothesis, where circRNAs were found associated with exon skipping events (Kelly et al. 2015). We are exploring this using the

datasets produced in this work, in collaboration with Petar Glažar and Nikolaus Rajewsky. Understanding whether and which circRNAs may modulate gene expression or protein function should have far reaching implications in our understanding of gene regulation during development and in disease.

To conclude, this work provides novel insights on gene expression regulation and brings the fields of transcription and circRNA biology closer together. Results presented here raise exciting questions and lay a solid ground for future explorations. Finally, this work highlights the importance of interdisciplinary research to tackle meaningful questions in the biology of complex systems and in disease.

Part V

Bibliography

6 Bibliography

Abranches, E., M. Silva, L. Pradier, H. Schulz, O. Hummel, D. Henrique, and E. Bekman. 2009. 'Neural differentiation of embryonic stem cells in vitro: a road map to neurogenesis in the embryo', *PLoS One*, 4: e6286.

Adelman, K., M. A. Kennedy, S. Nechaev, D. A. Gilchrist, G. W. Muse, Y. Chinenov, and I. Rogatsky. 2009. 'Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling', *Proc Natl Acad Sci U S A*, 106: 18207-12.

Adelman, K., and J. T. Lis. 2012. 'Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans', *Nat Rev Genet*, 13: 720-31.

Ahn, S. H., M. Kim, and S. Buratowski. 2004. 'Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing', *Mol Cell*, 13: 67-76.

Aida, M., Y. Chen, K. Nakajima, Y. Yamaguchi, T. Wada, and H. Handa. 2006. 'Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene', *Mol Cell Biol*, 26: 6094-104.

Aiyar, S. E., J. L. Sun, A. L. Blair, C. A. Moskaluk, Y. Z. Lu, Q. N. Ye, Y. Yamaguchi, A. Mukherjee, D. M. Ren, H. Handa, and R. Li. 2004. 'Attenuation of estrogen receptor alpha-mediated transcription through estrogen-stimulated recruitment of a negative elongation factor', *Genes Dev*, 18: 2134-46.

Akhtar, J., N. Kreim, F. Marini, G. Mohana, D. Brune, H. Binder, and J. Y. Roignant. 2019. 'Promoter-proximal pausing mediated by the exon junction complex regulates splicing', *Nat Commun*, 10: 521.

Akhtar, M. S., M. Heidemann, J. R. Tietjen, D. W. Zhang, R. D. Chapman, D. Eick, and A. Z. Ansari. 2009. 'TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II', *Mol Cell*, 34: 387-93.

Aktas, T., I. Avsar Ilik, D. Maticzka, V. Bhardwaj, C. Pessoa Rodrigues, G. Mittler, T. Manke, R. Backofen, and A. Akhtar. 2017. 'DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome', *Nature*, 544: 115-19.

Alexander, R. D., S. A. Innocente, J. D. Barrass, and J. D. Beggs. 2010. 'Splicing-dependent RNA polymerase pausing in yeast', *Mol Cell*, 40: 582-93.

Ali, I., D. G. Ruiz, Z. Ni, J. R. Johnson, H. Zhang, P. C. Li, M. M. Khalid, R. J. Conrad, X. Guo, J. Min, J. Greenblatt, M. Jacobson, N. J. Krogan, and M. Ott. 2019. 'Crosstalk between RNA Pol II C-Terminal Domain Acetylation and Phosphorylation via RPRD Proteins', *Mol Cell*, 74: 1164-74.e4.

Allen, B. L., and D. J. Taatjes. 2015. 'The Mediator complex: a central integrator of transcription', *Nat Rev Mol Cell Biol*, 16: 155-66.

Ameur, A., A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk. 2011. 'Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain', *Nat Struct Mol Biol*, 18: 1435-40.

An, D., R. Fujiki, D. E. Iannitelli, J. W. Smerdon, S. Maity, M. F. Rose, A. Gelber, E. K. Wanaselja, I. Yagudayeva, J. Y. Lee, C. Vogel, H. Wichterle, E. C. Engle, and E. O. Mazzoni. 2019. 'Stem cell-derived cranial and spinal motor neurons reveal proteostatic differences between ALS resistant and sensitive motor neurons', *Elife*, 8.

Andrews, S. 2010. 'FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.'.

Ashwal-Fluss, R., M. Meyer, N. R. Pamudurti, A. Ivanov, O. Bartok, M. Hanan, N. Evantal, S. Memczak, N. Rajewsky, and S. Kadener. 2014. 'circRNA biogenesis competes with pre-mRNA splicing', *Mol Cell*, 56: 55-66.

Aslanzadeh, V., Y. Huang, G. Sanguinetti, and J. D. Beggs. 2018. 'Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast', *Genome Res*, 28: 203-13.

Aufiero, S., M. M. G. van den Hoogenhof, Y. J. Reckman, A. Beqqali, I. van der Made, J. Kluin, M. A. F. Khan, Y. M. Pinto, and E. E. Creemers. 2018. 'Cardiac circRNAs arise mainly from constitutive exons rather than alternatively spliced exons', *Rna*, 24: 815-27.

Bachmayr-Heyda, Anna, Agnes T Reiner, Katharina Auer, Nyamdelger Sukhbaatar, Stefanie Aust, Thomas Bachleitner-Hofmann, Ildiko Mesteri, Thomas W Grunt, Robert Zeillinger, and Dietmar Pils. 2015. 'Correlation of circular RNA abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues', *Scientific reports*, 5: 8057.

Baralle, F. E., and J. Giudice. 2017. 'Alternative splicing as a regulator of development and tissue identity', *Nat Rev Mol Cell Biol*, 18: 437-51.

Barbosa-Morais, N. L., M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. 2012. 'The evolutionary landscape of alternative splicing in vertebrate species', *Science*, 338: 1587-93.

Bartman, C. R., N. Hamagami, C. A. Keller, B. Giardine, R. C. Hardison, G. A. Blobel, and A. Raj. 2019. 'Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation', *Mol Cell*, 73: 519-32.e4.

Bartolomei, M. S., N. F. Halden, C. R. Cullen, and J. L. Corden. 1988. 'Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II', *Mol Cell Biol*, 8: 330-9.

Baskaran, R., G. G. Chiang, and J. Y. Wang. 1996. 'Identification of a binding site in c-Ab1 tyrosine kinase for the C-terminal repeated domain of RNA polymerase II', *Mol Cell Biol*, 16: 3361-9.

Baskaran, R., M. E. Dahmus, and J. Y. Wang. 1993. 'Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain', *Proc Natl Acad Sci U S A*, 90: 11167-71.

Batsche, E., M. Yaniv, and C. Muchardt. 2006. 'The human SWI/SNF subunit Brm is a regulator of alternative splicing', *Nat Struct Mol Biol*, 13: 22-9.

Berget, S. M. 1995. 'Exon recognition in vertebrate splicing', *J Biol Chem*, 270: 2411-4.

Bernecky, C., J. M. Plitzko, and P. Cramer. 2017. 'Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp', *Nat Struct Mol Biol*, 24: 809-15.

Bisgrove, D. A., T. Mahmoudi, P. Henklein, and E. Verdin. 2007. 'Conserved P-TEFb-interacting domain of BRD4 inhibits HIV transcription', *Proc Natl Acad Sci U S A*, 104: 13690-5.

Bosken, C. A., L. Farnung, C. Hintermair, M. Merzel Schachter, K. Vogel-Bachmayr, D. Blazek, K. Anand, R. P. Fisher, D. Eick, and M. Geyer. 2014. 'The structure and substrate specificity of human Cdk12/Cyclin K', *Nat Commun*, 5: 3505.

Braun, J. E., L. J. Friedman, J. Gelles, and M. J. Moore. 2018. 'Synergistic assembly of human pre-spliceosomes across introns and exons', *Elife*, 7.

Braunschweig, U., S. Gueroussov, A. M. Plocik, B. R. Graveley, and B. J. Blencowe. 2013. 'Dynamic integration of splicing within gene regulatory pathways', *Cell*, 152: 1252-69.

Brookes, E., I. de Santiago, D. Hebenstreit, K. J. Morris, T. Carroll, S. Q. Xie, J. K. Stock, M. Heidemann, D. Eick, N. Nozaki, H. Kimura, J. Ragoussis, S. A. Teichmann, and A. Pombo. 2012. 'Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs', *Cell Stem Cell*, 10: 157-70.

Brookes, E., and A. Pombo. 2009. 'Modifications of RNA polymerase II are pivotal in regulating gene expression states', *EMBO Rep*, 10: 1213-9.

Capel, B., A. Swain, S. Nicolis, A. Hacker, M. Walter, P. Koopman, P. Goodfellow, and R. Lovell-Badge. 1993. 'Circular transcripts of the testis-determining gene Sry in adult mouse testis', *Cell*, 73: 1019-30.

Carrozza, M. J., B. Li, L. Florens, T. Suganuma, S. K. Swanson, K. K. Lee, W. J. Shia, S. Anderson, J. Yates, M. P. Washburn, and J. L. Workman. 2005. 'Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription', *Cell*, 123: 581-92.

Chao, S. H., K. Fujinaga, J. E. Marion, R. Taube, E. A. Sausville, A. M. Senderowicz, B. M. Peterlin, and D. H. Price. 2000. 'Flavopiridol inhibits P-TEFb and blocks HIV-1 replication', *J Biol Chem*, 275: 28345-8.

Chapman, R. D., M. Heidemann, T. K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. 2007. 'Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7', *Science*, 318: 1780-2.

Chathoth, K. T., J. D. Barrass, S. Webb, and J. D. Beggs. 2014. 'A splicing-dependent transcriptional checkpoint associated with prespliceosome formation', *Mol Cell*, 53: 779-90.

Chen, F. X., E. R. Smith, and A. Shilatifard. 2018. 'Born to run: control of transcription elongation by RNA polymerase II', *Nat Rev Mol Cell Biol*, 19: 464-78.

Chen, F. X., A. R. Woodfin, A. Gardini, R. A. Rickels, S. A. Marshall, E. R. Smith, R. Shiekhattar, and A. Shilatifard. 2015. 'PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II', *Cell*, 162: 1003-15.

Chen, F. X., P. Xie, C. K. Collings, K. Cao, Y. Aoi, S. A. Marshall, E. J. Rendleman, M. Ugarenko, P. A. Ozark, A. Zhang, R. Shiekhattar, E. R. Smith, M. Q. Zhang, and A. Shilatifard. 2017. 'PAF1 regulation of promoter-proximal pause release via enhancer activation', *Science*, 357: 1294-98.

Chen, L. L. 2016. 'The biogenesis and emerging roles of circular RNAs', *Nat Rev Mol Cell Biol*, 17: 205-11.

Chen, S., V. Huang, X. Xu, J. Livingstone, F. Soares, J. Jeon, Y. Zeng, J. T. Hua, J. Petricca, H. Guo, M. Wang, F. Yousif, Y. Zhang, N. Donmez, M. Ahmed, S. Volik, A. Lapuk, M. L. K. Chua, L. E. Heisler, A. Foucal, N. S. Fox, M. Fraser, V. Bhandari, Y. J. Shiah, J. Guan, J. Li, M. Orain, V. Picard, H. Hovington, A. Bergeron, L. Lacombe, Y. Fradet, B. Tetu, S. Liu, F. Feng, X. Wu, Y. W. Shao, M. A. Komor, C. Sahinalp, C. Collins, Y. Hoogstrate, M. de Jong, R. J. A. Fijneman, T. Fei, G. Jenster, T. van der Kwast, R. G. Bristow, P. C. Boutros, and H. H. He. 2019. 'Widespread and Functional RNA Circularization in Localized Prostate Cancer', *Cell*, 176: 831-43.e22.

Chen, Y. G., A. T. Satpathy, and H. Y. Chang. 2017. 'Gene regulation in the immune system by long noncoding RNAs', *Nat Immunol*, 18: 962-72.

Cheng, S-W Grace, Michael A Kuzyk, Annie Moradian, Taka-Aki Ichu, Vicky C-D Chang, Jerry F Tien, Sarah E Vollett, Malachi Griffith, Marco A Marra, and Gregg B Morin. 2012. 'Interaction of cyclin-dependent kinase 12/CrkRS with cyclin K1 is required for the phosphorylation of the C-terminal domain of RNA polymerase II', *Molecular and Cellular Biology*, 32: 4691-704.

Chinta, S. J., and J. K. Andersen. 2005. 'Dopaminergic neurons', *Int J Biochem Cell Biol*, 37: 942-6.

Cho, E. J., M. S. Kobor, M. Kim, J. Greenblatt, and S. Buratowski. 2001. 'Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain', *Genes Dev*, 15: 3319-29.

Cho, E. J., T. Takagi, C. R. Moore, and S. Buratowski. 1997. 'mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain', *Genes Dev*, 11: 3319-26.

Cocquerelle, C., P. Daubersies, M. A. Majerus, J. P. Kerckaert, and B. Bailleul. 1992. 'Splicing with inverted order of exons occurs proximal to large introns', *EMBO J*, 11: 1095-8.

Comer, F. I., and G. W. Hart. 2001. 'Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II', *Biochemistry*, 40: 7845-52.

Conn, S. J., K. A. Pillman, J. Toubia, V. M. Conn, M. Salmanidis, C. A. Phillips, S. Roslan, A. W. Schreiber, P. A. Gregory, and G. J. Goodall. 2015. 'The RNA binding protein quaking regulates formation of circRNAs', *Cell*, 160: 1125-34.

Core, L. J., J. J. Waterfall, and J. T. Lis. 2008. 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters', *Science*, 322: 1845-8.

Cortazar, M. A., R. M. Sheridan, B. Erickson, N. Fong, K. Glover-Cutter, K. Brannan, and D. L. Bentley. 2019. 'Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism', *Mol Cell*, 76: 896-908.e4.

Coulon, A., M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow, and D. R. Larson. 2014. 'Kinetic competition during the transcription cycle results in stochastic RNA processing', *Elife*, 3.

Cramer, P., J. F. Caceres, D. Cazalla, S. Kadener, A. F. Muro, F. E. Baralle, and A. R. Kornblihtt. 1999. 'Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer', *Mol Cell*, 4: 251-8.

Cramer, P., C. G. Pesce, F. E. Baralle, and A. R. Kornblihtt. 1997. 'Functional association between promoter structure and transcript alternative splicing', *Proc Natl Acad Sci U S A*, 94: 11456-60.

Czudnochowski, N., C. A. Bosken, and M. Geyer. 2012. 'Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition', *Nat Commun*, 3: 842.

Das, R., J. Yu, Z. Zhang, M. P. Gygi, A. R. Krainer, S. P. Gygi, and R. Reed. 2007. 'SR proteins function in coupling RNAP II transcription to pre-mRNA splicing', *Mol Cell*, 26: 867-81.

David, C. J., A. R. Boyne, S. R. Millhouse, and J. L. Manley. 2011. 'The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex', *Genes Dev*, 25: 972-83.

Davis-Dusenbery, B. N., L. A. Williams, J. R. Klim, and K. Eggan. 2014. 'How to make spinal motor neurons', *Development*, 141: 491-501.

Day, D. S., B. Zhang, S. M. Stevens, F. Ferrari, E. N. Larschan, P. J. Park, and W. T. Pu. 2016. 'Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types', *Genome Biol*, 17: 120.

de la Mata, M., C. R. Alonso, S. Kadener, J. P. Fededa, M. Blaustein, F. Pelisch, P. Cramer, D. Bentley, and A. R. Kornblihtt. 2003. 'A slow RNA polymerase II affects alternative splicing in vivo', *Mol Cell*, 12: 525-32.

de la Mata, M., and A. R. Kornblihtt. 2006. 'RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20', *Nat Struct Mol Biol*, 13: 973-80.

DeLaney, E., and D. S. Luse. 2016. 'Gdown1 Associates Efficiently with RNA Polymerase II after Promoter Clearance and Displaces TFIIF during Transcript Elongation', *PLoS One*, 11: e0163649.

Descostes, N., M. Heidemann, L. Spinelli, R. Schuller, M. A. Maqbool, R. Fenouil, F. Koch, C. Innocenti, M. Gut, I. Gut, D. Eick, and J. C. Andrau. 2014. 'Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells', *Elife*, 3: e02105.

Devaiah, B. N., C. Case-Borden, A. Gegonne, C. H. Hsu, Q. Chen, D. Meerzaman, A. Dey, K. Ozato, and D. S. Singer. 2016. 'BRD4 is a histone acetyltransferase that evicts nucleosomes from chromatin', *Nat Struct Mol Biol*, 23: 540-8.

Devaiah, B. N., B. A. Lewis, N. Cherman, M. C. Hewitt, B. K. Albrecht, P. G. Robey, K. Ozato, R. J. Sims, 3rd, and D. S. Singer. 2012. 'BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain', *Proc Natl Acad Sci U S A*, 109: 6927-32.

Dey, A., J. Ellenberg, A. Farina, A. E. Coleman, T. Maruyama, S. Sciortino, J. Lippincott-Schwartz, and K. Ozato. 2000. 'A bromodomain protein, MCAP, associates with mitotic chromosomes and affects G(2)-to-M transition', *Mol Cell Biol*, 20: 6537-49.

Di Vona, C., D. Bezdan, A. B. Islam, E. Salichs, N. Lopez-Bigas, S. Ossowski, and S. de la Luna. 2015. 'Chromatin-wide profiling of DYRK1A reveals a role as a gene-specific RNA polymerase II CTD kinase', *Mol Cell*, 57: 506-20.

Dias, J. D., T. Rito, E. Torlai Triglia, A. Kukalev, C. Ferrai, M. Chotalia, E. Brookes, H. Kimura, and A. Pombo. 2015. 'Methylation of RNA polymerase II non-consensus Lysine residues marks early transcription in mammalian cells', *Elife*, 4.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Drexler, H. L., K. Choquet, and L. S. Churchman. 2020. 'Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores', *Mol Cell*, 77: 985-98.e8.

Drexler, Heather L, Karine Choquet, and L Stirling Churchman. 2019. 'Human co-transcriptional splicing kinetics and coordination revealed by direct nascent RNA sequencing', *bioRxiv*: 611020.

Dujardin, G., C. Lafaille, M. de la Mata, L. E. Marasco, M. J. Munoz, C. Le Jossic-Corcos, L. Corcos, and A. R. Kornblihtt. 2014. 'How slow RNA polymerase II elongation favors alternative exon skipping', *Mol Cell*, 54: 683-90.

Dye, M. J., N. Gromak, and N. J. Proudfoot. 2006. 'Exon tethering in transcription by RNA polymerase II', *Mol Cell*, 21: 849-59.

Egloff, S., H. Al-Rawaf, D. O'Reilly, and S. Murphy. 2009. 'Chromatin structure is implicated in "late" elongation checkpoints on the U2 snRNA and beta-actin genes', *Mol Cell Biol*, 29: 4002-13.

Egloff, S., D. O'Reilly, R. D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. 2007. 'Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression', *Science*, 318: 1777-9.

Egloff, S., J. Zaborowska, C. Laitem, T. Kiss, and S. Murphy. 2012. 'Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes', *Mol Cell*, 45: 111-22.

Ehara, H., T. Yokoyama, H. Shigematsu, S. Yokoyama, M. Shirouzu, and S. I. Sekine. 2017. 'Structure of the complete elongation complex of RNA polymerase II with basal factors', *Science*, 357: 921-24.

Ehrensberger, A. H., G. P. Kelly, and J. Q. Svejstrup. 2013. 'Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps', *Cell*, 154: 713-5.

Eick, D., and M. Geyer. 2013. 'The RNA polymerase II carboxy-terminal domain (CTD) code', *Chem Rev*, 113: 8456-90.

Enuka, Y., M. Lauriola, M. E. Feldman, A. Sas-Chen, I. Ulitsky, and Y. Yarden. 2016. 'Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor', *Nucleic Acids Res*, 44: 1370-83.

Errichelli, L., S. Dini Modigliani, P. Laneve, A. Colantoni, I. Legnini, D. Capauto, A. Rosa, R. De Santis, R. Scarfo, G. Peruzzi, L. Lu, E. Caffarelli, N. A. Shneider, M. Morlando, and I. Bozzoni. 2017. 'FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons', *Nat Commun*, 8: 14741.

Fabrega, C., V. Shen, S. Shuman, and C. D. Lima. 2003. 'Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II', *Mol Cell*, 11: 1549-61.

Fei, T., Y. Chen, T. Xiao, W. Li, L. Cato, P. Zhang, M. B. Cotter, M. Bowden, R. T. Lis, S. G. Zhao, Q. Wu, F. Y. Feng, M. Loda, H. H. He, X. S. Liu, and M. Brown. 2017. 'Genome-wide CRISPR screen identifies HNRNPL as a prostate cancer dependency regulating RNA splicing', *Proc Natl Acad Sci U S A*, 114: E5207-e15.

Ferrai, C., E. Torlai Triglia, J. R. Risner-Janiczek, T. Rito, O. J. Rackham, I. de Santiago, A. Kukalev, M. Nicodemi, A. Akalin, M. Li, M. A. Ungless, and A. Pombo. 2017. 'RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation', *Mol Syst Biol*, 13: 946.

Floyd, S. R., M. E. Pacold, Q. Huang, S. M. Clarke, F. C. Lam, I. G. Cannell, B. D. Bryson, J. Rameseder, M. J. Lee, E. J. Blake, A. Fydrych, R. Ho, B. A. Greenberger, G. C. Chen, A. Maffa, A. M. Del Rosario, D. E. Root, A. E. Carpenter, W. C. Hahn, D. M. Sabatini, C. C. Chen, F. M. White, J. E. Bradner, and M. B. Yaffe. 2013. 'The bromodomain protein Brd4 insulates chromatin from DNA damage signalling', *Nature*, 498: 246-50.

Fong, N., H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu, and D. L. Bentley. 2014. 'Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate', *Genes Dev*, 28: 2663-76.

Fraser, J., C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken, S. Q. Xie, K. J. Morris, M. Itoh, H. Kawaji, I. Jaeger, Y. Hayashizaki, P. Carninci, A. R. Forrest, C. A. Semple, J. Dostie, A. Pombo, and M. Nicodemi. 2015. 'Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation', *Mol Syst Biol*, 11: 852.

Ganem, C., F. Devaux, C. Torchet, C. Jacq, S. Quevillon-Cheruel, G. Labesse, C. Facca, and G. Faye. 2003. 'Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast', *EMBO J*, 22: 1588-98.

Gao, Y., J. Wang, and F. Zhao. 2015. 'CIRI: an efficient and unbiased algorithm for de novo circular RNA identification', *Genome Biol*, 16: 4.

Ghosh, A., S. Shuman, and C. D. Lima. 2011. 'Structural insights to how mammalian capping enzyme reads the CTD code', *Mol Cell*, 43: 299-310.

Gilchrist, D. A., S. Nechaev, C. Lee, S. K. Ghosh, J. B. Collins, L. Li, D. S. Gilmour, and K. Adelman. 2008. 'NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly', *Genes Dev*, 22: 1921-33.

Glover-Cutter, K., S. Larochelle, B. Erickson, C. Zhang, K. Shokat, R. P. Fisher, and D. L. Bentley. 2009. 'TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II', *Mol Cell Biol*, 29: 5455-64.

Gornemann, J., K. M. Kotovic, K. Hujer, and K. M. Neugebauer. 2005. 'Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex', *Mol Cell*, 19: 53-63.

Greifenberg, A. K., D. Honig, K. Pilarova, R. Duster, K. Bartholomeeusen, C. A. Bosken, K. Anand, D. Blazek, and M. Geyer. 2016. 'Structural and Functional Analysis of the Cdk13/Cyclin K Complex', *Cell Rep*, 14: 320-31.

Gressel, S., B. Schwalb, T. M. Decker, W. Qin, H. Leonhardt, D. Eick, and P. Cramer. 2017. 'CDK9-dependent RNA polymerase II pausing controls transcription initiation', *Elife*, 6.

Grunberg, S., and S. Hahn. 2013. 'Structural insights into transcription initiation by RNA polymerase II', *Trends Biochem Sci*, 38: 603-11.

Gruner, H., M. Cortes-Lopez, D. A. Cooper, M. Bauer, and P. Miura. 2016. 'CircRNA accumulation in the aging mouse brain', *Sci Rep*, 6: 38907.

Gu, B., D. Eick, and O. Bensaude. 2013. 'CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo', *Nucleic Acids Res*, 41: 1591-603.

Guo, J. U., V. Agarwal, H. Guo, and D. P. Bartel. 2014. 'Expanded identification and characterization of mammalian circular RNAs', *Genome Biol*, 15: 409.

Hampsey, M., and D. Reinberg. 2003. 'Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation', *Cell*, 113: 429-32.

Hansen, T. B., T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems. 2013. 'Natural RNA circles function as efficient microRNA sponges', *Nature*, 495: 384-8.

Hansen, T. B., E. D. Wiklund, J. B. Bramsen, S. B. Villadsen, A. L. Statham, S. J. Clark, and J. Kjems. 2011. 'miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA', *EMBO J*, 30: 4414-22.

Harlen, K. M., K. L. Trotta, E. E. Smith, M. M. Mosaheb, S. M. Fuchs, and L. S. Churchman. 2016. 'Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue', *Cell Rep*, 15: 2147-58.

Henriques, T., D. A. Gilchrist, S. Nechaev, M. Bern, G. W. Muse, A. Burkholder, D. C. Fargo, and K. Adelman. 2013. 'Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals', *Mol Cell*, 52: 517-28.

Herzel, L., D. S. M. Ottoz, T. Alpert, and K. M. Neugebauer. 2017. 'Splicing and transcription touch base: co-transcriptional spliceosome assembly and function', *Nat Rev Mol Cell Biol*, 18: 637-50.

Herzel, L., K. Straube, and K. M. Neugebauer. 2018. 'Long-read sequencing of nascent RNA reveals coupling among RNA processing events', *Genome Res*, 28: 1008-19.

Hintermair, C., M. Heidemann, F. Koch, N. Descostes, M. Gut, I. Gut, R. Fenouil, P. Ferrier, A. Flatley, E. Kremmer, R. D. Chapman, J. C. Andrau, and D. Eick. 2012. 'Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation', *EMBO J*, 31: 2784-97.

Hintermair, C., K. Voss, I. Forne, M. Heidemann, A. Flatley, E. Kremmer, A. Imhof, and D. Eick. 2016. 'Specific threonine-4 phosphorylation and function of RNA polymerase II CTD during M phase progression', *Sci Rep*, 6: 27401.

Hollander, D., S. Naftelberg, G. Lev-Maor, A. R. Kornblihtt, and G. Ast. 2016. 'How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?', *Trends Genet*, 32: 596-606.

Hsin, J. P., A. Sheth, and J. L. Manley. 2011. 'RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing', *Science*, 334: 683-6.

Hsu, P. L., F. Yang, W. Smith-Kinnaman, W. Yang, J. E. Song, A. L. Mosley, and G. Varani. 2014. 'Rtr1 is a dual specificity phosphatase that dephosphorylates Tyr1 and Ser5 on the RNA polymerase II CTD', *J Mol Biol*, 426: 2970-81.

Huang, Y., W. Li, X. Yao, Q. J. Lin, J. W. Yin, Y. Liang, M. Heiner, B. Tian, J. Hui, and G. Wang. 2012. 'Mediator complex regulates alternative mRNA processing via the MED23 subunit', *Mol Cell*, 45: 459-69.

Ip, J. Y., D. Schmidt, Q. Pan, A. K. Ramani, A. G. Fraser, D. T. Odom, and B. J. Blencowe. 2011. 'Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation', *Genome Res*, 21: 390-401.

Irimia, M., R. J. Weatheritt, J. D. Ellis, N. N. Parikshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallieres, J. Tapial, B. Raj, D. O'Hanlon, M. Barrios-Rodiles, M. J. Sternberg, S. P. Cordes, F. P. Roth, J. L. Wrana, D. H. Geschwind, and B. J. Blencowe. 2014. 'A highly conserved program of neuronal microexons is misregulated in autistic brains', *Cell*, 159: 1511-23.

Ivanov, A., S. Memczak, E. Wyler, F. Torti, H. T. Porath, M. R. Orejuela, M. Piechotta, E. Y. Levanon, M. Landthaler, C. Dieterich, and N. Rajewsky. 2015. 'Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals', *Cell Rep*, 10: 170-7.

Jaeger, I., C. Arber, J. R. Risner-Janiczek, J. Kuechler, D. Pritzsche, I. C. Chen, T. Naveenan, M. A. Ungless, and M. Li. 2011. 'Temporally controlled modulation of FGF/ERK signaling directs midbrain dopaminergic neural progenitor fate in mouse and human pluripotent stem cells', *Development*, 138: 4363-74.

Jang, M. K., K. Mochizuki, M. Zhou, H. S. Jeong, J. N. Brady, and K. Ozato. 2005. 'The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription', *Mol Cell*, 19: 523-34.

Jeck, W. R., J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff, and N. E. Sharpless. 2013. 'Circular RNAs are abundant, conserved, and associated with ALU repeats', *Rna*, 19: 141-57.

Jeronimo, C., A. R. Bataille, and F. Robert. 2013. 'The writers, readers, and functions of the RNA polymerase II C-terminal domain code', *Chem Rev*, 113: 8491-522.

Jeronimo, C., and F. Robert. 2017. 'The Mediator Complex: At the Nexus of RNA Polymerase II Transcription', *Trends Cell Biol*, 27: 765-83.

Ji, X., Y. Zhou, S. Pandit, J. Huang, H. Li, C. Y. Lin, R. Xiao, C. B. Burge, and X. D. Fu. 2013. 'SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase', *Cell*, 153: 855-68.

Jishage, M., S. Malik, U. Wagner, B. Uberheide, Y. Ishihama, X. Hu, B. T. Chait, A. Gnatt, B. Ren, and R. G. Roeder. 2012. 'Transcriptional regulation by Pol II(G) involving mediator and competitive interactions of Gdown1 and TFIIF with Pol II', *Mol Cell*, 45: 51-63.

Jishage, M., X. Yu, Y. Shi, S. J. Ganesan, W. Y. Chen, A. Sali, B. T. Chait, F. J. Asturias, and R. G. Roeder. 2018. 'Architecture of Pol II(G) and molecular mechanism of transcription regulation by Gdown1', *Nat Struct Mol Biol*, 25: 859-67.

Jonkers, I., H. Kwak, and J. T. Lis. 2014. 'Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons', *Elife*, 3: e02407.

Kadener, S., P. Cramer, G. Nogues, D. Cazalla, M. de la Mata, J. P. Fededa, S. E. Werbajh, A. Srebrow, and A. R. Kornblihtt. 2001. 'Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing', *EMBO J*, 20: 5759-68.

Karagiannis, J., and M. K. Balasubramanian. 2007. 'A cyclin-dependent kinase that promotes cytokinesis through modulating phosphorylation of the carboxy terminal domain of the RNA Pol II Rpb1p sub-unit', *PLoS One*, 2: e433.

Kelly, S., C. Greenman, P. R. Cook, and A. Papantonis. 2015. 'Exon Skipping Is Correlated with Exon Circularization', *J Mol Biol*, 427: 2414-17.

Kelly, W. G., M. E. Dahmus, and G. W. Hart. 1993. 'RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc', *J Biol Chem*, 268: 10416-24.

Khatter, H., M. K. Vorländer, and C. W. Müller. 2017. 'RNA polymerase I and III: similar yet unique', *Curr Opin Struct Biol*, 47: 88-94.

Khodor, Y. L., J. S. Menet, M. Tolan, and M. Rosbash. 2012. 'Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse', *Rna*, 18: 2174-86.

Kim, M., H. Suh, E. J. Cho, and S. Buratowski. 2009. 'Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7', *J Biol Chem*, 284: 26421-6.

Kim, T., and S. Buratowski. 2009. 'Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions', *Cell*, 137: 259-72.

Kininis, M., B. S. Chen, A. G. Diehl, G. D. Isaacs, T. Zhang, A. C. Siepel, A. G. Clark, and W. L. Kraus. 2007. 'Genomic analyses of transcription factor binding, histone acetylation, and gene expression reveal mechanistically distinct classes of estrogen-regulated promoters', *Mol Cell Biol*, 27: 5090-104.

Koga, M., M. Hayashi, and D. Kaida. 2015. 'Splicing inhibition decreases phosphorylation level of Ser2 in Pol II CTD', *Nucleic Acids Res*, 43: 8258-67.

Kolde, Raivo. 2015. "pheatmap: Pretty Heatmaps. R package version 1.0. 8." In.

Komarnitsky, P., E. J. Cho, and S. Buratowski. 2000. 'Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription', *Genes Dev*, 14: 2452-60.

Kornblihtt, A. R., I. E. Schor, M. Allo, G. Dujardin, E. Petrillo, and M. J. Munoz. 2013. 'Alternative splicing: a pivotal step between eukaryotic transcription and translation', *Nat Rev Mol Cell Biol*, 14: 153-65.

Kramer, M. C., D. Liang, D. C. Tatomer, B. Gold, Z. M. March, S. Cherry, and J. E. Wilusz. 2015. 'Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins', *Genes Dev*, 29: 2168-82.

Krishnamurthy, S., X. He, M. Reyes-Reyes, C. Moore, and M. Hampsey. 2004. 'Ssu72 Is an RNA polymerase II CTD phosphatase', *Mol Cell*, 14: 387-94.

Laitem, C., J. Zaborowska, M. Tellier, Y. Yamaguchi, Q. Cao, S. Egloff, H. Handa, and S. Murphy. 2015. 'CTCF regulates NELF, DSIF and P-TEFb recruitment during transcription', *Transcription*, 6: 79-90.

Langmead, B., and S. L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9: 357-9.

Larochelle, S., R. Amat, K. Glover-Cutter, M. Sanso, C. Zhang, J. J. Allen, K. M. Shokat, D. L. Bentley, and R. P. Fisher. 2012. 'Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II', *Nat Struct Mol Biol*, 19: 1108-15.

Legnini, I., G. Di Timoteo, F. Rossi, M. Morlando, F. Briganti, O. Sthandier, A. Fatica, T. Santini, A. Andronache, M. Wade, P. Laneve, N. Rajewsky, and I. Bozzoni. 2017. 'Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis', *Mol Cell*, 66: 22-37.e9.

Lens, Z., F. X. Cantrelle, R. Peruzzini, X. Hanoulle, F. Dewitte, E. Ferreira, J. L. Baert, D. Monte, M. Aumercier, V. Villeret, A. Verger, and I. Landrieu. 2017. 'Solution Structure of the N-Terminal Domain of Mediator Subunit MED26 and Molecular Characterization of Its Interaction with EAF1 and TAF7', *J Mol Biol*, 429: 3043-55.

Lewis, B. A., A. L. Burlingame, and S. A. Myers. 2016. 'Human RNA Polymerase II Promoter Recruitment in Vitro Is Regulated by O-Linked N-Acetylglucosaminyltransferase (OGT)', *J Biol Chem*, 291: 14056-61.

Lewis, J. D., E. Izaurralde, A. Jarmolowski, C. McGuigan, and I. W. Mattaj. 1996. 'A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site', *Genes Dev*, 10: 1683-98.

Li, B., and C. N. Dewey. 2011. 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics*, 12: 323.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

Li, H., Z. Zhang, B. Wang, J. Zhang, Y. Zhao, and Y. Jin. 2007. 'Wwp2-mediated ubiquitination of the RNA polymerase II large subunit in mouse embryonic pluripotent stem cells', *Mol Cell Biol*, 27: 5296-305.

Li, J., Y. Liu, H. S. Rhee, S. K. Ghosh, L. Bai, B. F. Pugh, and D. S. Gilmour. 2013. 'Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing', *Mol Cell*, 50: 711-22.

Li, Q., S. Zheng, A. Han, C. H. Lin, P. Stoilov, X. D. Fu, and D. L. Black. 2014. 'The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation', *Elife*, 3: e01201.

Li, X., C. X. Liu, W. Xue, Y. Zhang, S. Jiang, Q. F. Yin, J. Wei, R. W. Yao, L. Yang, and L. L. Chen. 2017. 'Coordinated circRNA Biogenesis and Function with NF90/NF110 in Viral Infection', *Mol Cell*, 67: 214-27.e7.

Li, X., L. Yang, and L. L. Chen. 2018. 'The Biogenesis, Functions, and Challenges of Circular RNAs', *Mol Cell*, 71: 428-42.

Li, Z., C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai, P. Zhu, Z. Chang, Q. Wu, Y. Zhao, Y. Jia, P. Xu, H. Liu, and G. Shan. 2015. 'Exon-intron circular RNAs regulate transcription in the nucleus', *Nat Struct Mol Biol*, 22: 256-64.

Liang, D., D. C. Tatomer, Z. Luo, H. Wu, L. Yang, L. L. Chen, S. Cherry, and J. E. Wilusz. 2017. 'The Output of Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is Limiting', *Mol Cell*, 68: 940-54.e3.

Liang, D., and J. E. Wilusz. 2014. 'Short intronic repeat sequences facilitate circular RNA production', *Genes Dev*, 28: 2233-47.

Licatalosi, D. D., M. Yano, J. J. Fak, A. Mele, S. E. Grabinski, C. Zhang, and R. B. Darnell. 2012. 'Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain', *Genes Dev*, 26: 1626-42.

Lin, C., A. S. Garrett, B. De Kumar, E. R. Smith, M. Gogol, C. Seidel, R. Krumlauf, and A. Shilatifard. 2011. 'Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC)', *Genes Dev*, 25: 1486-98.

Lin, S., G. Coutinho-Mansfield, D. Wang, S. Pandit, and X. D. Fu. 2008. 'The splicing factor SC35 has an active role in transcriptional elongation', *Nat Struct Mol Biol*, 15: 819-26.

Liu, Z., D. R. Scannell, M. B. Eisen, and R. Tjian. 2011. 'Control of embryonic stem cell lineage commitment by core promoter factor, TAF3', *Cell*, 146: 720-31.

Loven, J., H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young. 2013. 'Selective inhibition of tumor oncogenes by disruption of super-enhancers', *Cell*, 153: 320-34.

Loya, T. J., and D. Reines. 2016. 'Recent advances in understanding transcription termination by RNA polymerase II', *F1000Res*, 5.

Luo, Z., C. Lin, E. Guest, A. S. Garrett, N. Mohaghegh, S. Swanson, S. Marshall, L. Florens, M. P. Washburn, and A. Shilatifard. 2012. 'The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output', *Mol Cell Biol*, 32: 2608-17.

Luo, Z., C. Lin, and A. Shilatifard. 2012. 'The super elongation complex (SEC) family in transcriptional control', *Nat Rev Mol Cell Biol*, 13: 543-7.

Maass, P. G., P. Glazar, S. Memczak, G. Dittmar, I. Hollfinger, L. Schreyer, A. V. Sauer, O. Toka, A. Aiuti, F. C. Luft, and N. Rajewsky. 2017. 'A map of human circular RNAs in clinically relevant tissues', *J Mol Med (Berl)*, 95: 1179-89.

Makeyev, E. V., J. Zhang, M. A. Carrasco, and T. Maniatis. 2007. 'The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing', *Mol Cell*, 27: 435-48.

Mandal, S. S., C. Chu, T. Wada, H. Handa, A. J. Shatkin, and D. Reinberg. 2004. 'Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II', *Proc Natl Acad Sci U S A*, 101: 7572-7.

Martin, R. M., J. Rino, C. Carvalho, T. Kirchhausen, and M. Carmo-Fonseca. 2013. 'Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity', *Cell Rep*, 4: 1144-55.

Maslon, M. M., U. Braunschweig, S. Aitken, A. R. Mann, F. Kilanowski, C. J. Hunter, B. J. Blencowe, A. R. Kornblihtt, I. R. Adams, and J. F. Caceres. 2019. 'A slow transcription rate causes embryonic lethality and perturbs kinetic coupling of neuronal genes', *EMBO J*, 38.

Matera, A. G., and Z. Wang. 2014. 'A day in the life of the spliceosome', *Nat Rev Mol Cell Biol*, 15: 108-21.

Maudlin, I. E., and J. D. Beggs. 2019. 'Spt5 modulates cotranscriptional spliceosome assembly in Saccharomyces cerevisiae', *Rna*, 25: 1298-310.

Mavrich, T. N., C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh. 2008. 'Nucleosome organization in the Drosophila genome', *Nature*, 453: 358-62.

Maxon, M. E., J. A. Goodrich, and R. Tjian. 1994. 'Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance', *Genes Dev*, 8: 515-24.

Mayer, A., J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J. A. Stamatoyannopoulos, and L. S. Churchman. 2015. 'Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution', *Cell*, 161: 541-54.

Mayer, A., M. Heidemann, M. Lidschreiber, A. Schreieck, M. Sun, C. Hintermair, E. Kremmer, D. Eick, and P. Cramer. 2012. 'CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II', *Science*, 336: 1723-5.

Mazzoni, E. O., S. Mahony, M. Closser, C. A. Morrison, S. Nedelec, D. J. Williams, D. An, D. K. Gifford, and H. Wichterle. 2013. 'Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity', *Nat Neurosci*, 16: 1219-27.

McCracken, S., N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Foster, S. Shuman, and D. L. Bentley. 1997. '5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II', *Genes Dev*, 11: 3306-18.

McCracken, S., N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens, and D. L. Bentley. 1997. 'The C-terminal domain of RNA polymerase II couples mRNA processing to transcription', *Nature*, 385: 357-61.

McNamara, R. P., J. E. Reeder, E. A. McMillan, C. W. Bacon, J. L. McCann, and I. D'Orso. 2016. 'KAP1 Recruitment of the 7SK snRNP Complex to Promoters Enables Transcription Elongation by RNA Polymerase II', *Mol Cell*, 61: 39-53.

Meinhart, A., and P. Cramer. 2004. 'Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors', *Nature*, 430: 223-6.

Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, and N. Rajewsky. 2013. 'Circular RNAs are a large class of animal RNAs with regulatory potency', *Nature*, 495: 333-8.

Michels, A. A., V. T. Nguyen, A. Fraldi, V. Labas, M. Edwards, F. Bonnet, L. Lania, and O. Bensaude. 2003. 'MAQ1 and 7SK RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner', *Mol Cell Biol*, 23: 4859-69.

Milligan, L., C. Sayou, A. Tuck, T. Auchynnikava, J. E. Reid, R. Alexander, F. L. Alves, R. Allshire, C. Spanos, J. Rappsilber, J. D. Beggs, G. Kudla, and D. Tollervey. 2017. 'RNA polymerase II stalling at pre-mRNA splice sites is enforced by ubiquitination of the catalytic subunit', *Elife*, 6.

Morris, D. P., and A. L. Greenleaf. 2000. 'The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II', *J Biol Chem*, 275: 39935-43.

Mosley, A. L., S. G. Pattenden, M. Carey, S. Venkatesh, J. M. Gilmore, L. Florens, J. L. Workman, and M. P. Washburn. 2009. 'Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation', *Mol Cell*, 34: 168-78.

Munoz, M. J., M. S. Perez Santangelo, M. P. Paronetto, M. de la Mata, F. Pelisch, S. Boireau, K. Glover-Cutter, C. Ben-Dov, M. Blaustein, J. J. Lozano, G. Bird, D. Bentley, E. Bertrand, and A. R. Kornblihtt. 2009. 'DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation', *Cell*, 137: 708-20.

Myers, L. C., C. M. Gustafsson, D. A. Bushnell, M. Lui, H. Erdjument-Bromage, P. Tempst, and R. D. Kornberg. 1998. 'The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain', *Genes Dev*, 12: 45-54.

Narita, T., T. M. Yung, J. Yamamoto, Y. Tsuboi, H. Tanabe, K. Tanaka, Y. Yamaguchi, and H. Handa. 2007. 'NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs', *Mol Cell*, 26: 349-65.

Nechaev, S., and K. Adelman. 2011. 'Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation', *Biochim Biophys Acta*, 1809: 34-45.

Ng, H. H., F. Robert, R. A. Young, and K. Struhl. 2003. 'Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity', *Mol Cell*, 11: 709-19.

Nguyen, V. T., T. Kiss, A. A. Michels, and O. Bensaude. 2001. '7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes', *Nature*, 414: 322-5.

Ni, Z., A. Saunders, N. J. Fuda, J. Yao, J. R. Suarez, W. W. Webb, and J. T. Lis. 2008. 'P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo', *Mol Cell Biol*, 28: 1161-70.

Nigro, J. M., K. R. Cho, E. R. Fearon, S. E. Kern, J. M. Ruppert, J. D. Oliner, K. W. Kinzler, and B. Vogelstein. 1991. 'Scrambled exons', *Cell*, 64: 607-13.

Nogues, G., S. Kadener, P. Cramer, D. Bentley, and A. R. Kornblihtt. 2002. 'Transcriptional activators differ in their abilities to control alternative splicing', *J Biol Chem*, 277: 43110-4.

Nojima, T., T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot. 2015. 'Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing', *Cell*, 161: 526-40.

Nojima, T., K. Rebelo, T. Gomes, A. R. Grosso, N. J. Proudfoot, and M. Carmo-Fonseca. 2018. 'RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing', *Mol Cell*, 72: 369-79.e4.

Oesterreich, F. C., L. Herzel, K. Straube, K. Hujer, J. Howard, and K. M. Neugebauer. 2016. 'Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II', *Cell*, 165: 372-81.

Ozsolak, F., J. S. Song, X. S. Liu, and D. E. Fisher. 2007. 'High-throughput mapping of the chromatin structure of human promoters', *Nat Biotechnol*, 25: 244-8.

Pabis, M., N. Neufeld, M. C. Steiner, T. Bojic, Y. Shav-Tal, and K. M. Neugebauer. 2013. 'The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells', *Rna*, 19: 1054-63.

Pai, A. A., T. Henriques, K. McCue, A. Burkholder, K. Adelman, and C. B. Burge. 2017. 'The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture', *Elife*, 6.

Pak, V., T. T. Eifler, S. Jager, N. J. Krogan, K. Fujinaga, and B. M. Peterlin. 2015. 'CDK11 in TREX/THOC Regulates HIV mRNA 3' End Processing', *Cell Host Microbe*, 18: 560-70.

Pamudurti, N. R., O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, S. Shenzis, M. Samson, G. Dittmar, M. Landthaler, M. Chekulaeva, N. Rajewsky, and S. Kadener. 2017. 'Translation of CircRNAs', *Mol Cell*, 66: 9-21.e7.

Parua, P. K., G. T. Booth, M. Sansó, B. Benjamin, J. C. Tanny, J. T. Lis, and R. P. Fisher. 2018. 'A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II', *Nature*, 558: 460-64.

Patop, I. L., and S. Kadener. 2018. 'circRNAs in Cancer', *Curr Opin Genet Dev*, 48: 121-27.

Pei, Y., and S. Shuman. 2002. 'Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5', *J Biol Chem*, 277: 19639-48.

Peng, J., Y. Zhu, J. T. Milton, and D. H. Price. 1998. 'Identification of multiple cyclin subunits of human P-TEFb', *Genes Dev*, 12: 755-62.

Peterlin, B. M., and D. H. Price. 2006. 'Controlling the elongation phase of transcription with P-TEFb', *Mol Cell*, 23: 297-305.

Piwecka, M., P. Glazar, L. R. Hernandez-Miranda, S. Memczak, S. A. Wolf, A. Rybak-Wolf, A. Filipchyk, F. Klironomos, C. A. Cerda Jara, P. Fenske, T. Trimbuch, V. Zywitza, M. Plass, L. Schreyer, S. Ayoub, C. Kocks, R. Kuhn, C. Rosenmund, C. Birchmeier, and N. Rajewsky. 2017. 'Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function', *Science*, 357.

Porrua, O., and D. Libri. 2015. 'Transcription termination and the control of the transcriptome: why, where and how to stop', *Nat Rev Mol Cell Biol*, 16: 190-202.

Proudfoot, N. J. 2011. 'Ending the message: poly(A) signals then and now', *Genes Dev*, 25: 1770-82.

Qiu, Y., and D. S. Gilmour. 2017. 'Identification of Regions in the Spt5 Subunit of DRB Sensitivity-inducing Factor (DSIF) That Are Involved in Promoter-proximal Pausing', *J Biol Chem*, 292: 5555-70.

Rabani, M., R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit, and A. Regev. 2014. 'High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies', *Cell*, 159: 1698-710.

Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. 2010. 'c-Myc regulates transcriptional pause release', *Cell*, 141: 432-45.

Raj, B., M. Irimia, U. Braunschweig, T. Sterne-Weiler, D. O'Hanlon, Z. Y. Lin, G. I. Chen, L. E. Easton, J. Ule, A. C. Gingras, E. Eyras, and B. J. Blencowe. 2014. 'A global regulatory mechanism for activating an exon network required for neurogenesis', *Mol Cell*, 56: 90-103.

Ranuncolo, Stella M, Salil Ghosh, John A Hanover, Gerald W Hart, and Brian A Lewis. 2012. 'Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo', *Journal of Biological Chemistry*, 287: 23549-61.

Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. 'Exon definition may facilitate splice site selection in RNAs with multiple exons', *Mol Cell Biol*, 10: 84-94.

Rosonina, E., and B. J. Blencowe. 2004. 'Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3'-end cleavage', *Rna*, 10: 581-9.

Ryan, K., K. G. Murthy, S. Kaneko, and J. L. Manley. 2002. 'Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3' cleavage', *Mol Cell Biol*, 22: 1684-92.

Rybak-Wolf, A., C. Stottmeister, P. Glazar, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, M. Herzog, L. Schreyer, P. Papavasileiou, A. Ivanov, M. Ohman, D. Refojo, S. Kadener, and N. Rajewsky. 2015. 'Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed', *Mol Cell*, 58: 870-85.

Sainsbury, S., C. Bernecky, and P. Cramer. 2015. 'Structural basis of transcription initiation by RNA polymerase II', *Nat Rev Mol Cell Biol*, 16: 129-43.

Salzman, J. 2016. 'Circular RNA Expression: Its Potential Regulation and Function', *Trends Genet*, 32: 309-16.

Salzman, J., R. E. Chen, M. N. Olsen, P. L. Wang, and P. O. Brown. 2013. 'Cell-type specific features of circular RNA expression', *PLoS Genet*, 9: e1003777.

Salzman, J., C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown. 2012. 'Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types', *PLoS One*, 7: e30733.

Sanger, H. L., G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt. 1976. 'Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures', *Proc Natl Acad Sci U S A*, 73: 3852-6.

Schaaf, C. A., H. Kwak, A. Koenig, Z. Misulovin, D. W. Gohara, A. Watson, Y. Zhou, J. T. Lis, and D. Dorsett. 2013. 'Genome-wide control of RNA polymerase II activity by cohesin', *PLoS Genet*, 9: e1003382.

Schneider, M., C. L. Will, M. Anokhina, J. Tazi, H. Urlaub, and R. Luhrmann. 2010. 'Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes', *Mol Cell*, 38: 223-35.

Schones, D. E., K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. 2008. 'Dynamic regulation of nucleosome positioning in the human genome', *Cell*, 132: 887-98.

Schreieck, A., A. D. Easter, S. Etzold, K. Wiederhold, M. Lidschreiber, P. Cramer, and L. A. Passmore. 2014. 'RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7', *Nat Struct Mol Biol*, 21: 175-79.

Schroder, S., E. Herker, F. Itzen, D. He, S. Thomas, D. A. Gilchrist, K. Kaehlcke, S. Cho, K. S. Pollard, J. A. Capra, M. Schnolzer, P. A. Cole, M. Geyer, B. G. Bruneau, K. Adelman, and M. Ott. 2013. 'Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells', *Mol Cell*, 52: 314-24.

Schuller, R., I. Forne, T. Straub, A. Schreieck, Y. Texier, N. Shah, T. M. Decker, P. Cramer, A. Imhof, and D. Eick. 2016. 'Heptad-Specific Phosphorylation of RNA Polymerase II CTD', *Mol Cell*, 61: 305-14.

Schwartz, J. C., C. C. Ebmeier, E. R. Podell, J. Heimiller, D. J. Taatjes, and T. R. Cech. 2012. 'FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2', *Genes Dev*, 26: 2690-5.

Schwer, B., and S. Shuman. 2011. 'Deciphering the RNA polymerase II CTD code in fission yeast', *Mol Cell*, 43: 311-8.

Shao, W., and J. Zeitlinger. 2017. 'Paused RNA polymerase II inhibits new transcriptional initiation', *Nat Genet*, 49: 1045-51.

Sharma, S., L. A. Kohlstaedt, A. Damianov, D. C. Rio, and D. L. Black. 2008. 'Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome', *Nat Struct Mol Biol*, 15: 183-91.

Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski, M. Kashlev, P. Oberdoerffer, R. Sandberg, and S. Oberdoerffer. 2011. 'CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing', *Nature*, 479: 74-9.

Sims, R. J., 3rd, R. Belotserkovskaya, and D. Reinberg. 2004. 'Elongation by RNA polymerase II: the short and long of it', *Genes Dev*, 18: 2437-68.

Singh, J., and R. A. Padgett. 2009. 'Rates of in situ transcription and splicing in large human genes', *Nat Struct Mol Biol*, 16: 1128-33.

Skourti-Stathaki, K., and N. J. Proudfoot. 2014. 'A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression', *Genes Dev*, 28: 1384-96.

Skourti-Stathaki, K., N. J. Proudfoot, and N. Gromak. 2011. 'Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination', *Mol Cell*, 42: 794-805.

Sogaard, T. M., and J. Q. Svejstrup. 2007. 'Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator', *J Biol Chem*, 282: 14113-20.

St Amour, C. V., M. Sanso, C. A. Bosken, K. M. Lee, S. Larochelle, C. Zhang, K. M. Shokat, M. Geyer, and R. P. Fisher. 2012. 'Separate domains of fission yeast Cdk9 (P-TEFb) are required for capping enzyme recruitment and primed (Ser7-phosphorylated) Rpb1 carboxyl-terminal domain substrate recognition', *Mol Cell Biol*, 32: 2372-83.

Stadelmayer, B., G. Micas, A. Gamot, P. Martin, N. Malirat, S. Koval, R. Raffel, B. Sobhian, D. Severac, S. Rialle, H. Parrinello, O. Cuvier, and M. Benkirane. 2014. 'Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes', *Nat Commun*, 5: 5531.

Starke, S., I. Jost, O. Rossbach, T. Schneider, S. Schreiner, L. H. Hung, and A. Bindereif. 2015. 'Exon circularization requires canonical splice signals', *Cell Rep*, 10: 103-11.

Stock, J. K., S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A. G. Fisher, and A. Pombo. 2007. 'Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells', *Nat Cell Biol*, 9: 1428-35.

Suh, H., S. B. Ficarro, U. B. Kang, Y. Chun, J. A. Marto, and S. Buratowski. 2016. 'Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II', *Mol Cell*, 61: 297-304.

Sun, J., and R. Li. 2010. 'Human negative elongation factor activates transcription and regulates alternative transcription initiation', *J Biol Chem*, 285: 6443-52.

Takahashi, H., T. J. Parmely, S. Sato, C. Tomomori-Sato, C. A. Banks, S. E. Kong, H. Szutorisz, S. K. Swanson, S. Martin-Brown, M. P. Washburn, L. Florens, C. W. Seidel, C. Lin, E. R. Smith, A. Shilatifard, R. C. Conaway, and J. W. Conaway. 2011. 'Human mediator subunit MED26 functions as a docking site for transcription elongation factors', *Cell*, 146: 92-104.

Tee, W. W., S. S. Shen, O. Oksuz, V. Narendra, and D. Reinberg. 2014. 'Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs', *Cell*, 156: 678-90.

Thompson, C. M., A. J. Koleske, D. M. Chao, and R. A. Young. 1993. 'A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast', *Cell*, 73: 1361-75.

Tietjen, J. R., D. W. Zhang, J. B. Rodriguez-Molina, B. E. White, M. S. Akhtar, M. Heidemann, X. Li, R. D. Chapman, K. Shokat, S. Keles, D. Eick, and A. Z. Ansari. 2010. 'Chemical-genomic dissection of the CTD code', *Nat Struct Mol Biol*, 17: 1154-61.

Tilgner, H., D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigo. 2012. 'Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs', *Genome Res*, 22: 1616-25.

Vargas, D. Y., K. Shah, M. Batish, M. Levandoski, S. Sinha, S. A. Marras, P. Schedl, and S. Tyagi. 2011. 'Single-molecule imaging of transcriptionally coupled and uncoupled splicing', *Cell*, 147: 1054-65.

Velasco, S., M. M. Ibrahim, A. Kakumanu, G. Garipler, B. Aydin, M. A. Al-Sayegh, A. Hirsekorn, F. Abdul-Rahman, R. Satija, U. Ohler, S. Mahony, and E. O. Mazzoni. 2017. 'A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells', *Cell Stem Cell*, 20: 205-17.e8.

Veloso, A., K. S. Kirkconnell, B. Magnuson, B. Biewen, M. T. Paulsen, T. E. Wilson, and M. Ljungman. 2014. 'Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications', *Genome Res*, 24: 896-905.

Venkatesh, S., M. Smolle, H. Li, M. M. Gogol, M. Saint, S. Kumar, K. Natarajan, and J. L. Workman. 2012. 'Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes', *Nature*, 489: 452-5.

Viladevall, L., C. V. St Amour, A. Rosebrock, S. Schneider, C. Zhang, J. J. Allen, K. M. Shokat, B. Schwer, J. K. Leatherwood, and R. P. Fisher. 2009. 'TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast', *Mol Cell*, 33: 738-51.

Vos, S. M., L. Farnung, M. Boehning, C. Wigge, A. Linden, H. Urlaub, and P. Cramer. 2018. 'Structure of activated transcription complex Pol II-DSIF-PAF-SPT6', *Nature*, 560: 607-12.

Vos, S. M., D. Pollmann, L. Caizzi, K. B. Hofmann, P. Rombaut, T. Zimniak, F. Herzog, and P. Cramer. 2016. 'Architecture and RNA binding of the human negative elongation factor', *Elife*, 5.

Voss, K., I. Forne, N. Descostes, C. Hintermair, R. Schuller, M. A. Maqbool, M. Heidemann, A. Flatley, A. Imhof, M. Gut, I. Gut, E. Kremmer, J. C. Andrau, and D. Eick. 2015. 'Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II', *Transcription*, 6: 91-101.

Vuong, C. K., D. L. Black, and S. Zheng. 2016. 'The neurogenetics of alternative splicing', *Nat Rev Neurosci*, 17: 265-81.

Wen, Y., and A. J. Shatkin. 1999. 'Transcription elongation factor hSPT5 stimulates mRNA capping', *Genes Dev*, 13: 1774-9.

West, M. L., and J. L. Corden. 1995. 'Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations', *Genetics*, 140: 1223-33.

Westholm, J. O., P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, S. E. Celniker, B. R. Graveley, and E. C. Lai. 2014. 'Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation', *Cell Rep*, 9: 1966-80.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. 2013. 'Master transcription factors and mediator establish super-enhancers at key cell identity genes', *Cell*, 153: 307-19.

Wickham, Hadley. 2016. *ggplot2: elegant graphics for data analysis* (Springer).

Will, C. L., and R. Luhrmann. 2011. 'Spliceosome structure and function', *Cold Spring Harb Perspect Biol*, 3.

Wilusz, J. E. 2018. 'A 360 degrees view of circular RNAs: From biogenesis to functions', *Wiley Interdiscip Rev RNA*, 9: e1478.

Windhager, L., T. Bonfert, K. Burger, Z. Ruzsics, S. Krebs, S. Kaufmann, G. Malterer, A. L'Hernault, M. Schilhabel, S. Schreiber, P. Rosenstiel, R. Zimmer, D. Eick, C. C. Friedel, and L. Dolken. 2012. 'Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution', *Genome Res*, 22: 2031-42.

Wong, K. H., Y. Jin, and K. Struhl. 2014. 'TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape', *Mol Cell*, 54: 601-12.

Workman, J. L., and R. G. Roeder. 1987. 'Binding of transcription factor TFIID to the major late promoter during in vitro nucleosome assembly potentiates subsequent initiation by RNA polymerase II', *Cell*, 51: 613-22.

Wu, C. H., Y. Yamaguchi, L. R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. 2003. 'NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila', *Genes Dev*, 17: 1402-14.

Xing, H., Y. Mo, W. Liao, and M. Q. Zhang. 2012. 'Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data', *PLoS Comput Biol*, 8: e1002613.

Xu, Y., C. Bernecky, C. T. Lee, K. C. Maier, B. Schwalb, D. Tegunov, J. M. Plitzko, H. Urlaub, and P. Cramer. 2017. 'Architecture of the RNA polymerase II-Paf1C-TFIIS transcription elongation complex', *Nat Commun*, 8: 15741.

Yamaguchi, Y., N. Inukai, T. Narita, T. Wada, and H. Handa. 2002. 'Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA', *Mol Cell Biol*, 22: 2918-27.

Yamaguchi, Y., T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. 1999. 'NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation', *Cell*, 97: 41-51.

Yamamoto, J., Y. Hagiwara, K. Chiba, T. Isobe, T. Narita, H. Handa, and Y. Yamaguchi. 2014. 'DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes', *Nat Commun*, 5: 4263.

Yang, Y., X. Fan, M. Mao, X. Song, P. Wu, Y. Zhang, Y. Jin, Y. Yang, L. L. Chen, Y. Wang, C. C. Wong, X. Xiao, and Z. Wang. 2017. 'Extensive translation of circular RNAs driven by N(6)-methyladenosine', *Cell Res*, 27: 626-41.

Yang, Z., J. H. Yik, R. Chen, N. He, M. K. Jang, K. Ozato, and Q. Zhou. 2005. 'Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4', *Mol Cell*, 19: 535-45.

Yang, Z., Q. Zhu, K. Luo, and Q. Zhou. 2001. 'The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription', *Nature*, 414: 317-22.

Ying, Q. L., M. Stavridis, D. Griffiths, M. Li, and A. Smith. 2003. 'Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture', *Nat Biotechnol*, 21: 183-6.

You, X., I. Vlatkovic, A. Babic, T. Will, I. Epstein, G. Tushev, G. Akbalik, M. Wang, C. Glock, C. Quedenau, X. Wang, J. Hou, H. Liu, W. Sun, S. Sambandan, T. Chen, E. M. Schuman, and W. Chen. 2015. 'Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity', *Nat Neurosci*, 18: 603-10.

Yu, M., W. Yang, T. Ni, Z. Tang, T. Nakadai, J. Zhu, and R. G. Roeder. 2015. 'RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II', *Science*, 350: 1383-6.

Zaborowska, J., S. Egloff, and S. Murphy. 2016. 'The pol II CTD: new twists in the tail', *Nat Struct Mol Biol*, 23: 771-7.

Zeitlinger, J., A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. 2007. 'RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo', *Nat Genet*, 39: 1512-6.

Zhang, D. W., A. L. Mosley, S. R. Ramisetty, J. B. Rodriguez-Molina, M. P. Washburn, and A. Z. Ansari. 2012. 'Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination', *J Biol Chem*, 287: 8541-51.

Zhang, H., F. Rigo, and H. G. Martinson. 2015. 'Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change

Mechanism that Does Not Require Cleavage at the Poly(A) Site', *Mol Cell*, 59: 437-48.

Zhang, X. O., H. B. Wang, Y. Zhang, X. Lu, L. L. Chen, and L. Yang. 2014. 'Complementary sequence-mediated exon circularization', *Cell*, 159: 134-47.

Zhang, Y., W. Xue, X. Li, J. Zhang, S. Chen, J. L. Zhang, L. Yang, and L. L. Chen. 2016. 'The Biogenesis of Nascent Circular RNAs', *Cell Rep*, 15: 611-24.

Zhang, Y., X. O. Zhang, T. Chen, J. F. Xiang, Q. F. Yin, Y. H. Xing, S. Zhu, L. Yang, and L. L. Chen. 2013. 'Circular intronic long noncoding RNAs', *Mol Cell*, 51: 792-806.

Zhao, D. Y., G. Gish, U. Braunschweig, Y. Li, Z. Ni, F. W. Schmitges, G. Zhong, K. Liu, W. Li, J. Moffat, M. Vedadi, J. Min, T. J. Pawson, B. J. Blencowe, and J. F. Greenblatt. 2016. 'SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination', *Nature*, 529: 48-53.

Zhong, H., J. Zhu, H. Zhang, L. Ding, Y. Sun, C. Huang, and Q. Ye. 2004. 'COBRA1 inhibits AP-1 transcriptional activity in transfected cells', *Biochem Biophys Res Commun*, 325: 568-73.

Zylka, Mark J, Jeremy M Simon, and Benjamin D Philpot. 2015. 'Gene length matters in neurons', *Neuron*, 86: 353-55.

Part VI

Appendix

7 Appendix

Figure 1.2

Figure 1.2 is reproduced with permission (see below) from “Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans”, Nature Reviews Genetics, 2012

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4672641486741
License date	Sep 19, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans
Licensed Content Author	Karen Adelman et al
Licensed Content Date	Sep 18, 2012
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019
Portions	Figure 3

Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 1.4

Figure 1.4 is reproduced with permission (see below) from “A day in the life of the spliceosome”, Nature Reviews Molecular Cell Biology, 2014

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676641437838
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	A day in the life of the spliceosome
Licensed Content Author	A. Gregory Matera, Zefeng Wang
Licensed Content Date	Jan 23, 2014
Licensed Content Volume	15
Licensed Content Issue	2
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association

Expected presentation date	Oct 2019
Portions	Fig. 4
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 1.5 and 1.6

Figure 1.5 and 1.6 are reproduced with permission (see below) from “The neurogenetics of alternative splicing”, Nature Reviews Neuroscience, 2016.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676621452198
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Neuroscience
Licensed Content Title	The neurogenetics of alternative splicing
Licensed Content Author	Celine K. Vuong, Douglas L. Black, Sika Zheng
Licensed Content Date	Apr 20, 2016
Licensed Content Volume	17
Licensed Content Issue	5
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	3
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no

Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019
Portions	Box 1, box 2, fig. 1a
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 1.7

Figure 1.7 is reproduced with permission (see below) from “How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?”, Trends in Genetics, 2016.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4676630364363
License date	Sep 26, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Genetics
Licensed Content Title	How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?
Licensed Content Author	Dror Hollander,Shiran Naftelberg,Galit Lev-Maor,Alberto R. Kornblihtt,Gil Ast
Licensed Content Date	Oct 1, 2016
Licensed Content Volume	32
Licensed Content Issue	10
Licensed Content Pages	11
Start Page	596
End Page	606
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations

Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig. 2b
Title of your thesis/dissertation	Dynamic regulation of co-transcriptional processes during neuronal maturation
Publisher of new work	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected completion date	Oct 2019
Estimated size (number of pages)	1
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR

Figure 1.8 and 1.9

Figure 1.8 and 1.9 are reproduced with permission (see below) from “Alternative splicing: a pivotal step between eukaryotic transcription and translation”, Nature Reviews Molecular Cell Biology, 2013.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676630683686
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	Alternative splicing: a pivotal step between eukaryotic transcription and translation
Licensed Content Author	Alberto R. Kornblihtt et al
Licensed Content Date	Feb 6, 2013

Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019
Portions	Fig. 2 and 3
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 1.10

Figure 1.10 is reproduced with permission (see below) from “Circular RNA Expression: Its Potential Regulation and Function”, Trends in Genetics, 2016.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4676630858325
License date	Sep 26, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Genetics
Licensed Content Title	Circular RNA Expression: Its Potential Regulation and Function
Licensed Content Author	Julia Salzman

Licensed Content Date	May 1, 2016
Licensed Content Volume	32
Licensed Content Issue	5
Licensed Content Pages	8
Start Page	309
End Page	316
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig. 1
Title of your thesis/dissertation	Dynamic regulation of co-transcriptional processes during neuronal maturation
Publisher of new work	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected completion date	Oct 2019
Estimated size (number of pages)	1
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR

Figure 1.11 and 1.12

Figure 1.11 and 1.12 are reproduced with permission (see below) from “The biogenesis and emerging roles of circular RNAs”, Nature Reviews Molecular Cell Biology, 2016.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676631072010
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	The biogenesis and emerging roles of circular RNAs
Licensed Content Author	Ling-Ling Chen
Licensed Content Date	Feb 24, 2016
Licensed Content Volume	17
Licensed Content Issue	4
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019
Portions	Fig. 1
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 2.1

Figure 2.1 is reproduced with permission (see below) from “Circular RNAs are a large class of animal RNAs with regulatory potency”, Nature, 2013.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676640813126
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature
Licensed Content Title	Circular RNAs are a large class of animal RNAs with regulatory potency
Licensed Content Author	Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger et al.
Licensed Content Date	Feb 27, 2013
Licensed Content Volume	495
Licensed Content Issue	7441
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019
Portions	Fig. 1
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany

	Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR

Figure 3.3 A)

Figure 3.3 A) is reproduced with permission (see below) from "Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity", Nature Neuroscience, 2013.

This Agreement between BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association -- Ana Miguel Fernandes ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4676641193073
License date	Sep 26, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Neuroscience
Licensed Content Title	Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity
Licensed Content Author	Esteban O Mazzoni, Shaun Mahony, Michael Closser, Carolyn A Morrison, Stephane Nedelec et al.
Licensed Content Date	Jul 21, 2013
Licensed Content Volume	16
Licensed Content Issue	9
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Dynamic regulation of co-transcriptional processes during neuronal maturation
Institution name	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Expected presentation date	Oct 2019

Portions	Fig. 1
Requestor Location	BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association Hannoversche str. 28 Berlin, 10115 Germany Attn: BIMSB, Max Delbrück Center for Molecular Medicine in the Helmholtz Association
Total	0.00 EUR